UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

Programa de Maestría en Microbiología, Parasitología, Química Clínica e Inmunología

ASSESSMENT OF MICROBIAL COMMUNITIES ASSOCIATED WITH THE GUT AND FEEDING SUBSTRATE OF THREE XYLOPHAGOUS COLEOPTERAN FAMILIES WITH EMPHASIS ON CELLULOSE-DEGRADING BACTERIA

EVALUACIÓN DE LAS COMUNIDADES MICROBIANAS ASOCIADAS AL INTESTINO Y SUSTRATO ALIMENTICIO DE TRES FAMILIAS DE COLEÓPTEROS CON ÉNFASIS EN BACTERIAS DEGRADADORAS DE CELULOSA

Tesis sometida a la consideración de la Comisión del Programa de Posgrado en Microbiología, Parasitología, Química Clínica e Inmunología para optar al grado y título de Maestría Académica en Microbiología

IBRAHIM ZÚÑIGA CHAVES

Ciudad Universitaria Rodrigo Facio, Costa Rica

2019

"Esta Tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Microbiología, Parasitología, Química Clínica e Inmunología de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Microbiologia con Enfasis en Bacteriologia"

Carlos E. Rodríguez

PhD. Carlos Rodriguez Rodriguez

Decano o Representante del Decano

Sistema de Estudios de Posgrado

PhD. Adrian Pinto Tomas

Profesor Guía

Catalina

Msc. Catalina Murillo Cruz

Lectora

PhD. Cesar Rodriguez Sanchez

Lector

Rebeca

PhD. Rebeca Campos Sanchez

Director (a) Coordinador (a) /Representante

Programa de Posgrado en Posgrado en Microbiología, Parasitología, Química Clínica e Inmunología

Ibrahim Zuniga Chaves

Sustentante

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Ibrahim Zuniga Chaves, con cédula de identidad 113550263, en mi condición de autor del TFG titulado Evaluación de las comunidades microbianas asociadas al intestino y sustrato alimenticio de tres familias de coleópteros con énfasis en bacterias degradadoras de celulosa.

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** ✓ **NO** * ☐

***En caso de la negativa favor indicar el tiempo de restricción:** ______ **año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Ibrahim Zuñiga Chaves.

Número de Carné: A66388 Número de cédula: 113550263.

Correo Electrónico: ibrahim.zuniga@ucr.ac.cr.

Fecha: 27-11-2019. Número de teléfono: ______.

Nombre del Director (a) de Tesis o Tutor (a): Adrián Pinto Tomás.

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

Agradecimientos

A mi director de tesis, Adrian Pinto, por casi 9 años de ser un excelente mentor y guía. También gracias por su entusiasmo y motivación para perseguir una carrera en investigación.

A Catalina Murillo por siempre ser un apoyo y guiarme durante todo el proceso de mi posgrado.

A Cesar Rodríguez por el entrenamiento en bioinformática y su apoyo en la construcción de este trabajo.

A Gabriel Vargas por su mentoría en el mundo de la bioinformática y la ecología microbiana.

A la sección de bioprospección de INBio por toda la colecta de muestras y apoyo para el desarrollo de todo el trabajo bioinformatico.

A mis padres Flor Heydi Chaves y Jorge Arturo Zuñiga por que nunca han dejado de creer en mí y todo lo que he logrado es gracias a ellos.

Tabla de contenidos

Abstract

The current energy crisis and depletion of fossil fuel reserves has encouraged research for new alternatives. Degradation of plant material by enzymatic reactions represents an important strategy to obtain affordable and renewable energy. In nature, several beetle species have solved this problem by developing digestive systems colonized by specialized microbiota capable of breaking down recalcitrant carbon through cellulose-breaking enzymes, releasing sugars that later are used as nutrients. Therefore, their associated microbes represent potential sources of novel bioenergy-relevant molecules. Previous work, centered on one beetle Family (Passalidae) from one location in Costa Rica, revealed the occurrence of distinct microbial communities among larvae, adults and their woody substrate inside the same decomposing log. Furthermore, larval metagenomes were enriched in Firmicutes (particularly Ruminococcaceae) sequences and they harbor multiple genes coding for cellulose degrading enzymes. In the present study, I first evaluated whether the unique communities observed for Passalid beetles remain stable in other locations and compared them with other families of xylophagous beetles. This goal was achieved by sequencing 16S rRNA gene libraries from a total of 84 Coleopteran-associated samples collected in 3 different Costa Rican regions (Braulio Carrillo, Corcovado and Isla del Coco National Parks). These samples included the intestinal content of beetles (larvae and adults) belonging to 3 different families (Passalidae, Scarabaeidae and Cerambycidae) and the woody substrate they consume. Employing NMDS analyses, I confirmed that microbial communities are mainly driven by sample type (larvae, adult or substrate), but are also influenced by the other two variables: beetle family and geographic location. Also, I confirmed that Firmicutes is the most abundant phylum in both adult (40%) and larval (35%) samples from all beetles tested and that Ruminococcaceae (15%) is the most abundant family in larvae. The high abundance of Ruminococcaceae OTUs in larvae suggests an important role of these organisms in the overall gut metabolism and cellulose breakdown from the wood they feed on. To evaluate this possibility, available metagenomes and a collection of metagenome assembled genomes (MAGs) from Passalid beetles were employed to analyze the potential of the

Ruminococcaceae family to degrade cellulose. From the metagenomes, 2-3% of all genes related to glycosil hydrolases (GHs) were assigned to the Ruminococcaceae family. Furthermore, out of the 11 Ruminococcaceae MAGs assembled in this work, 9 did not cluster with any of the reference genomes included in the analysis. Four of them have putative cellulases, while two of them, Bin 519 and 174, have gene clusters with several components related to cellulosomes. The present work suggests that anaerobic bacteria, such as Ruminococcaceae, are dominant members of the microbial community in the digestive tracts of Scarabaeidae and Passalidae larvae. Further, these organisms seem to partake directly in cellulose degradation and they also intervene in other metabolic reactions. Considering that these MAGs represent potentially undescribed organisms that are both present in high abundance in the gut of these xylophagous beetles and encode for important genes necessary to obtain nutrients out of their recalcitrant food source, they are ideal candidates for further isolation in pure culture to elucidate their definitive role in this system.

Resumen

La crisis energética actual y el agotamiento de las reservas de combustibles fósiles han fomentado la investigación para obtener nuevas alternativas. La descomposición de material vegetal por reacciones enzimáticas constituye una oportunidad para obtener energía renovable de manera económica. En la naturaleza, varias especies de escarabajos han resuelto este problema al desarrollar un sistema digestivo colonizado por microbiota capaz de descomponer celulosa a través de enzimas especializadas, liberando azúcares que luego utilizan como suministro de nutrientes. Por lo tanto, estos microorganismos representan fuentes potenciales de nuevas moléculas para la investigación en bioenergía. Trabajos previos, enfocados en una sola familia de escarabajos (Passalidae), en una única ubicación geográfica en Costa Rica, evidenciaron la presencia de comunidades microbianas únicas y distintas entre larvas, adultos y el sustrato asociado dentro del mismo tronco en descomposición. Además, los metagenomas de las larvas están enriquecidos con secuencias de Firmicutes y genes que codifican por enzimas capaces de romper celulosa. En el presente estudio se empleó la secuenciación del gen que codifica para el ARNr 16S para evaluar la composición de la comunidad bacteriana de un total de 84 muestras asociadas a coleópteros recolectadas en 3 diferentes regiones de Costa Rica (Parques Nacionales Braulio Carrillo, Corcovado e Isla del Coco). Estas muestras incluyeron el contenido intestinal de escarabajos (larvas y adultos) pertenecientes a 3 familias diferentes (Passalidae, Scarabaeidae y Cerambycidae) y el sustrato leñoso que consumen. Al emplear un análisis de beta diversidad, se observó que la agrupación de las comunidades microbianas se explica principalmente por el tipo de muestra (larva, adulto o sustrato), aunque la familia del escarabajo y la ubicación geográfica también tiene una incidencia estadísticamente significativa en la composición de la comunidad. Este análisis confirmó que Firmicutes es el filo más abundante en muestras de adultos (40%) y larvas (35%) de todos los escarabajos evaluados, además Ruminococcaceae (15%) es la familia más abundante en las larvas. La gran abundancia de OTUs de Ruminococcaceae en este estadio sugiere un papel importante de estos organismos en el metabolismo general del intestino y su participación en la degradación de la celulosa presente en la madera de la que se alimentan. Para evaluar esta posibilidad, se utilizaron metagenomas disponibles de la

familia Passalidae y una colección de genomas ensamblados de metagenomas (MAGs) con el fin de analizar el potencial de la familia Ruminococcaceae para degradar celulosa. En los metagenomas, 2-3% de los genes relacionados con glicosil hidrolasas (GH) se asignaron a la familia Ruminococcaceae. Además, de los 11 MAGs ensamblados pertenecientes a la familia Ruminococcaceae, 9 no se agruparon con ninguno de los genomas de referencia incluidos en el análisis. Cuatro de estos MAGs tienen potenciales celulasas, y entre ellos, el Bin 519 y el Bin 174 tienen agrupaciones de genes relacionados con la estructura de celulosomas. Los resultados anteriores sugieren que bacterias anaerobias, como Ruminococcaceae, son miembros dominantes de la comunidad microbiana en los tractos digestivos de las larvas de la familia Scarabaeidae y Passalidae. Adicionalmente, estos microorganismos parecen tener participación en la degradación de celulosa e intervienen como intermediarios en otras reacciones metabólicas. Por lo tanto, dado que estos MAGs representan organismos presentes en gran abundancia en el intestino de estos escarabajos xilófagos y codifican por genes importantes para la obtención de nutrientes a partir de su substrato alimenticio, es recomendable obtenerlos en cultivo puro para elucidar su contribución en la ecofisiología de sus hospederos.

List of Tables

List of Figures

List of abbreviations

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

16S rRNA: Gen of the 16S ribosomal small subunit

CBM: Cellulose binding modules

GH: Glycosyl hydrolases

COG: Cluster of orthologous genes

HMM: Hidden Markov models

LEFSE: Linear discriminant analysis effect size

LDA: Linear discriminant analysis

MAG: Metagenome assembled genomes.

NMDS: Non metric distance scale

OTU: Operational taxonomic unit

SCFA: short chain fatty acids

NMDS: non metric distance scale

ICBG: International Cooperative Biodiversity Groups

INBio: National Institute of Biodiversity

IMG: Integrated Microbial Genomes

ACCVC: Central Volcanic Mountain Range Conservation Area

ACOSA: Osa Conservation Area

ACMIC: Cocos Island Conservation Area

SCCG: Single copy core gen

1. Introduction

1.1 Insect-Microorganisms symbiosis

Microorganisms, ubiquitous in the biosphere, play key roles in biogeochemistry cycles through specialized and unique metabolic pathways [1]. They can also be crucial in the evolutionary success of other living beings through the establishment of symbiotic associations [2,3]. Bacteria, fungi and protozoan genomes encode enzymes not present in their more complex hosts, allowing the gain of adaptive advantages over other animals competing in the same ecosystem [4,5]. Insects are a widely studied example of this type of symbiosis. Many species hold mutualistic associations with microorganisms that enhance arthropod defenses against pathogens[6] and cooperate in other essential functions for their survival including digestion, nitrogen fixation and vitamin and amino acid synthesis [7–9].

One of the best characterized associations occurs between aphids and the ɤ-Proteobacteria, *Buchnera aphidicola*, an endosymbiont that has developed a close mutualistic relationship with certain species of aphids that feed on plant sap [10]. *B. aphidicola* bacteria reside within specialized polyploid cells called bacteriocytes, where they collaborate with the synthesis of essential amino acids that the aphid would not be able to acquire otherwise [2]. Transmission of *B. aphidicola* occurs through vertical passage between adult aphids that inoculate their eggs with the bacteria, assuring its presence in their progeny [11]. Another example of a strict mutualism occurs between lower termites and their gut microbiota. Lower termite species are able to seize more nutrients from their Nitrogen-limited diet (rich in lignin and cellulose) due to their associations with eukaryote flagellates residing inside their gut [12].

1.2. Microbial cellulose degradation

A common trait in plants, fungi and bacterial cells is the existence of a cell wall. However, in each of these groups, its function and chemical composition is different. Plant cell walls are composed mostly by polysaccharides such as cellulose, lignin, hemicellulose, pectin and, in fewer proportion, by cutin, suberin, cerum and proteins [13]. Such compounds are distributed in two different layers: the first one is composed by cellulose, hemicellulose and pectin (Figure 1A). The second layer is a more rigid, insoluble and irregular network composed by lignin and polymeric phenylpropanoid subunits including p-hydroxyphenyl, guaiacil and siringil [14].

Lignin is the main component of tree bark and one of the most abundant plant residues on both agricultural and natural ecosystems [15]. On the other hand, cellulose (Figure 1B) is a carbon polymer composed by linear chains of several hundred to thousands of $\beta(1\rightarrow4)$ linked D-glucose units and constitutes the most abundant organic material on Earth [16]. Cellulose fibers (Figure 1C) line with each other in an anti-parallel arrangement, forming a sort of flat strip constituted by repetitive units of cellobiose (two D-glucose bound by β 1-4 bonds). Van der Waals forces bind the linear chains together and hydrogen bonds form the final linkage with the neighbor chains to create the microfibers of non-polar crystalline cellulose [14,17].

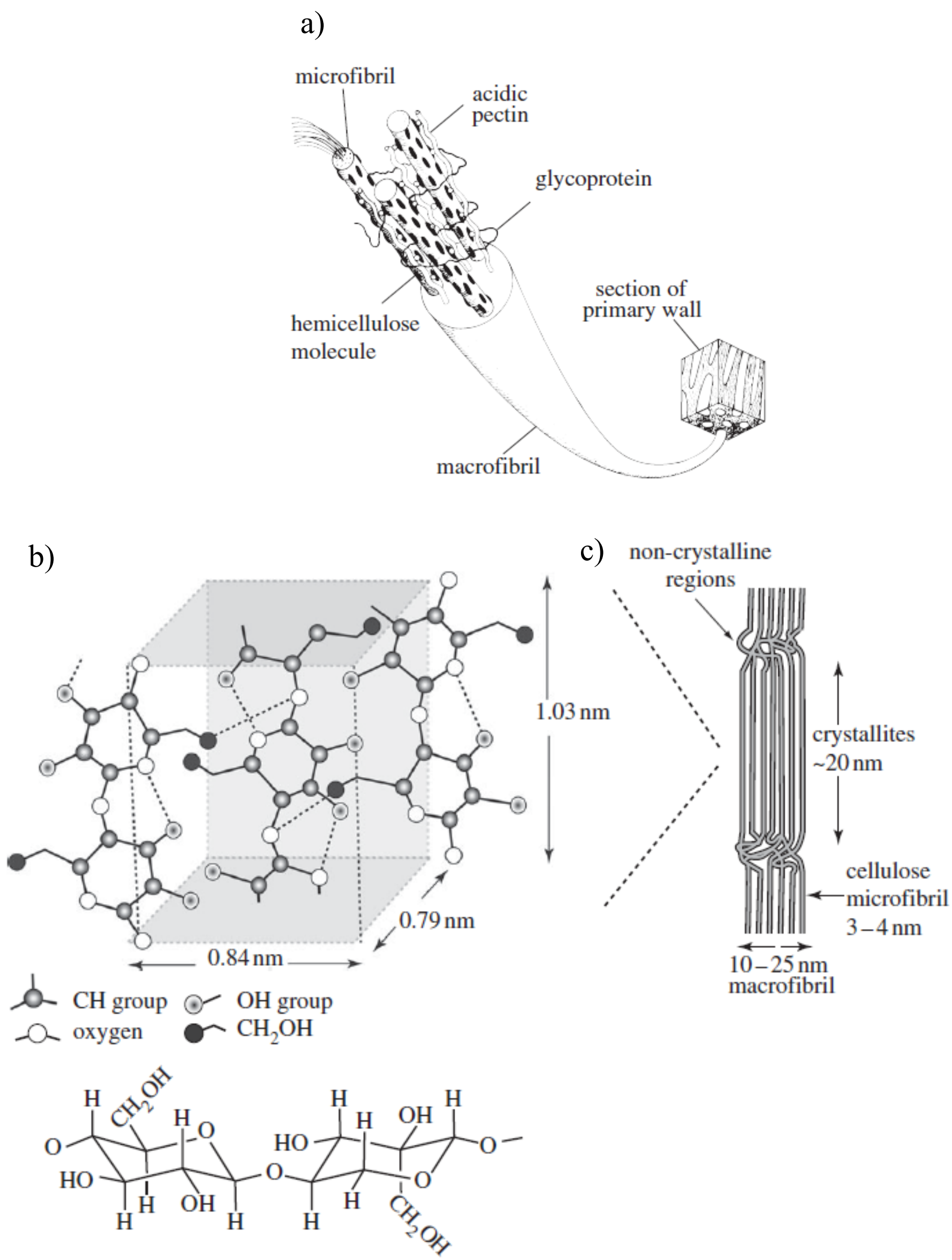


Figure 1. Plant cell wall hierarchic structure. a) Primary macrofibre b) Molecular structure of the cellulose carbohydrates. c) Cellulose microfibers (crystalline and non-crystalline). Modified from [14]

Another important component of the cell wall matrix is hemicellulose. This polymer shares similar traits with cellulose, however, it contains other carbohydrates besides glucose such as xylose, mannose, galactose, and arabinose. Hemicellulose chains are comprised by 500 to 3000 monosaccharides [16]. In general, cellulose content varies among plant species and oscillates between 35-50% of its dry weight [18], while hemicellulose represents 25-35%, lignin 12-35% and pectin 2-5% [14,19].

Cellulases comprise a wide group of glycosil hydrolases (proteins that break bonds in carbohydrates by adding a water molecule), including every enzyme capable to degrade or metabolize cellulose and associated polymers, except for lignin [20,21].Cellulases have either endocellulase or exocellulase activity or both, depending on which site(s) they operate to breakdown cellulose crystals [22].

Cellulases are classified in three main groups: endoglucanases (EC 3.2.1.4), exoglycanases (EC 3.2.1.74, EC 3.2.1.91) and β-glucosidases (EC 3.2.1.21)[23]. Endoglucanases hydrolyze β-1,4-glycoslidic bonds randomly on the inside of cellulose chains releasing oligosaccharides of different sizes. Exoglycanases execute their activity at reducing and non-reducing sites in cellulose crystals. Finally, β-glucosidases catalyze hydrolysis of non-reducing ends of cellobiose molecules to generate free glucose monosaccharides and do not carry out their function on the actual cellulose fibers [22].

Cellulases can also be secreted as part of multi-enzymatic complexes called cellulosomes [24]. Cellulosomes consist of a large macromolecular complex (40-180kDa) with different types of cellulases attached to scaffold proteins [24]. They are synthetized mostly by anaerobic bacteria such as *Clostridium sp.* and *Ruminococcus sp.*, but are also present in certain aerobic bacteria and fungi [20]. Cellulosomes rely on the synergistic activity of various enzymes, therefore are highly efficient degrading cellulose and other components of plant cell walls [25].

The structure of cellulosomes (Figure 2) are built around a "scaffolding" protein which may include membrane adhesion motifs, various cellulose binding modules (CBM) and several cohesin domains with the function of binding cellulolytic enzymes [26]. Cellulases in organisms that utilize cellulosomes contain repetitive sequences of approximately 20 amino acid residues called dockering domains. Dockering domains allow cellulases to attach to the cohesin domains present in the scaffolding protein [20].

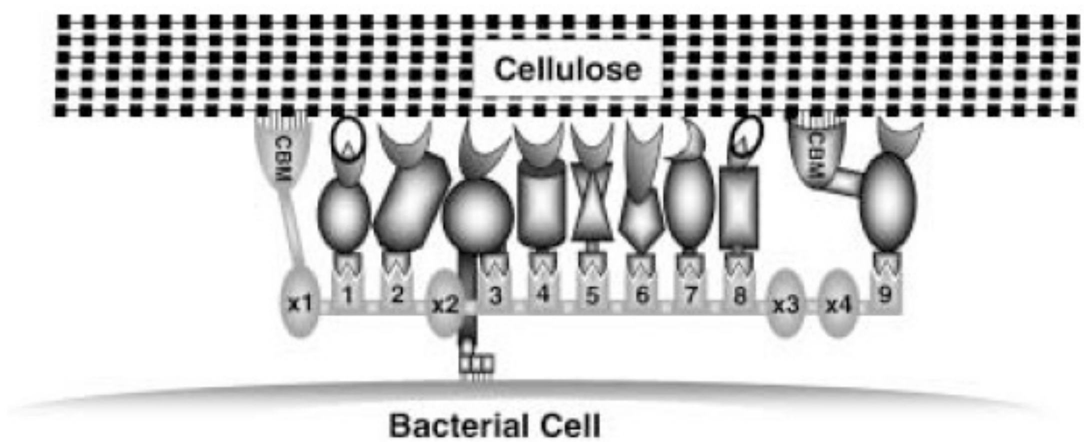


Figure 2. Hypothetic structure of a cellulosome. The scaffolding protein is drawn in light gray and the four small squares under it point the membrane binding motifs. All the other enzymatic components are represented in dark gray. Modified from [24].

1.3 Cellulose metabolism in microbe-host symbioses.

Microbial communities in association with herbivores facilitate the conversion of biomass from plants into nutrients that can be seized by the host [3]. Ruminants represent one of the most efficient systems of cellulose degradation studied to date [27]. Cattle and other ruminants host a specialized microbiota able to metabolize the plant material consumed by them (Figure 3). They ferment cellulose and other complex carbohydrates and turn them into glucose and short chain fatty acids (SCFA), later used as energy source [28]. Besides ruminants, other mammals such as sloths, horses and deer maintain symbiotic relationships with cellulose-degrading microorganisms inside their gut that contribute in the breakdown of the plant material they feed on [29–31].

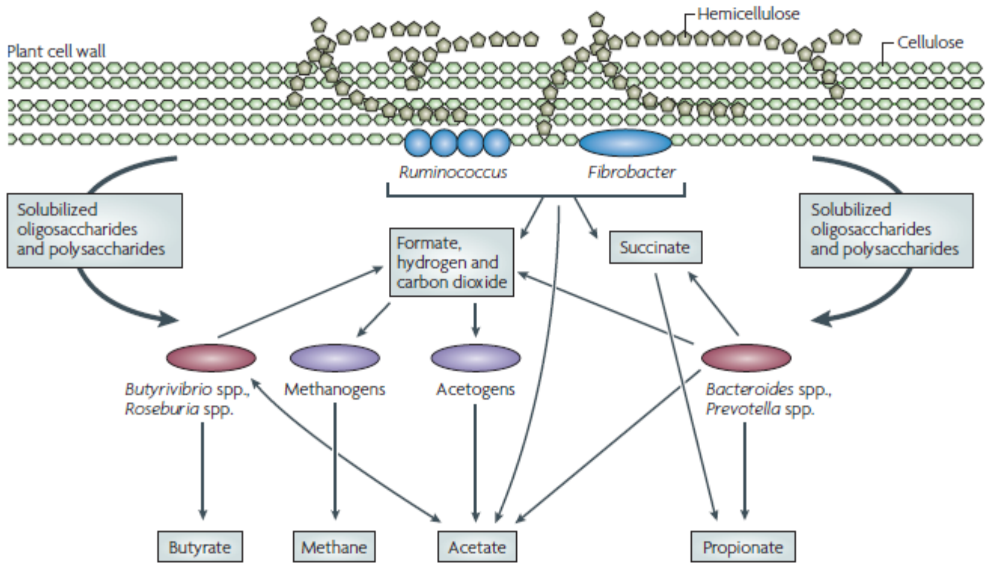


Figure 3. Cellulose and associated carbohydrate metabolism by ruminant gut microbiota. Modified from [28].

Mammals are not the only organisms capable of hosting cellulose degrading microbiota in their gut. Bacteria associated with xylophagous insects are capable of digesting cellulose in aerobic, microaerophilic and anaerobic conditions. Because of it, their hosts are capable of colonizing fallen logs and wooden structures they use as food source, or subsist by consuming large amounts of leaves [9,32]. For example, termites maintain colonies with hundreds of thousands individuals with wood as their sole food source thanks to well stablished symbiosis. Lower termites are colonized by flagellates from the phylum Parabasalia and higher termites rely on bacteria from the Proteobacteria and Spirochaetes phyla [12,33,34]. Another insect able to subsist feeding exclusively on wood are the Passalid beetles, a family of coleopterans with subsocial behavior. Passalid beetles have a strict xylophagous diet where all development stages, from larva to adult, subsist by consuming decomposing organic material from the fallen logs they inhabit.

1.4 Passalid beetle biology

Insects are the most diverse group of animals in nature. They have the capacity to inhabit a great diversity of ecosystems worldwide and thrive of almost every available source [35]. Amongst insects, coleopterans or beetles, as they are commonly known, represent the most diverse order [36]. From the total of 300 000 species of coleopterans estimated to exist, around 10% can be found in Costa Rica [37].

Coleopterans feed on a variety of diets, according to the family or subfamily they belong. The majority of beetles consume plant associated material, but there are also predator and fungivore beetles [37]. Finally, many species feed on decomposing organic matter (saproxylophagous) either from animal or plant sources (for example, fallen logs in the forest floor) [38].

The main families with saproxylophagous feeding behaviors include beetles from the Passalidae, Scarabaeidae, Tenebrionidae, Cerambycidae, Brentidae, Buprestidae, Curculionidae, Nitidulidae, Oedemeridae and Silphidae families [37]. In some families, such as Cerambycidae o Sacarabaeidae, larvae exhibit a xylophagous pattern, while adults might adopt different diets once they undergo their respective metamorphosis [39]. Members of the Passalidae family comprise a group with exclusive xylophagous behavior. Besides, both larvae and adults use the space formed under the tree bark as shelter to build their nest on fallen trees while devouring the portions starting to decompose.

The Passalidae family belongs to the suborder Polyphaga of the order Coleoptera, superfamily Scarabaeoidae, which in turn includes five subfamilies, among them only Passalinae and Proculinae are found in the Americas [38]. They are subsocial insects, which means they exhibit some type of parental care. All Passalid adults collaborate in feeding the offspring while coexisting with two or three other generations at the same time [40]. Passalid family groups start when a breeding couple meets and colonizes a fallen tree (Figure 4). Later, the female lay eggs and waits until they hatch, and then adults collaborate to provide an environment rich in woody substrate mixed with feces that larvae are capable of digesting. Once the larva is ready to become pupae, adults cover them with a mixture of semi-digested wood and fecal material. Consequently, this mixture is the first substrate they will eat once they complete their metamorphosis and become adults. Once the young adults emerge, they will help take care of the next generation until they become sexually active and leave the nest to colonize a new fallen tree [35].

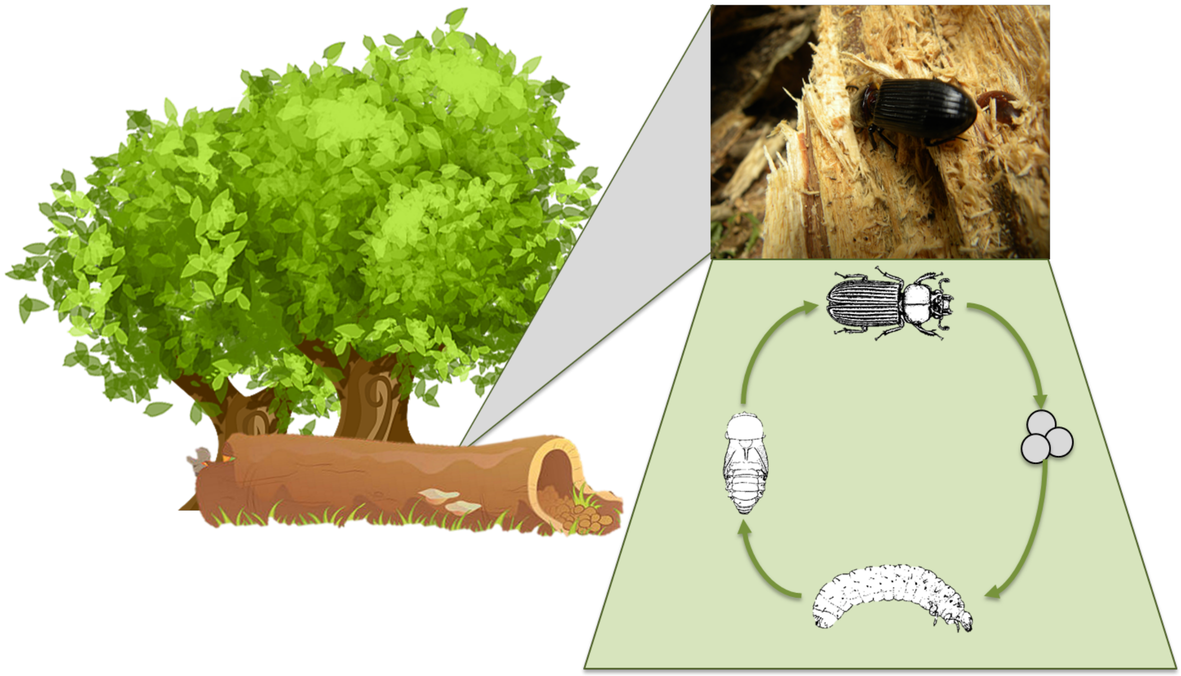


Figure 4. Life cycle of beetles from the Passalidae family. They exhibit a subsocial behavior with parental care that overlaps across generations.

In nature, the fallen trees Passalid beetles inhabit are usually very humid and have at least one month of decomposition [36]. Inside the tree, the insects build tunnels in the region between the bark and the actual wood, developing little tunnel-like spaces known as galleries [41]. Galleries are inhabited by all the individuals that form part of the family group. The adults feed the larvae with gallery material, which will also be called substrate, as part of their parental care. The substrate consists of a mixture of chewed and pre-digested wood and feces, similar to the mixture employed to cover the pupae [42]. This kind of behavior seems to indicate that constant feces and substrate consumption by larvae and even adults is imperative and could contribute to a constant exchange of microorganisms between all the members in one family group. In fact, Passalid larvae are not able to survive when deprived of parental care or when raised on sterilized substrate as exclusive food source, suggesting that the transmission of microorganisms through the gallery material is key for their survival [36,38].

1.5 Passalid beetle microbiota.

Insect digestive tracts, as most other similar ecosystems, represent confined environments that host a complex and diverse community of microorganisms, both transient and resident. They include organisms from the three domains of life: bacteria, archaea and microbial eukaryotes. The gut microbiota can reach a density of 10^9-10^{11} cells/ml and in some cases it is essential for various of their host functions such as development, nutrition and growth[43]. Additionally, the gut microbiota participates in pheromone production, pathogenesis and environmental adaptability processes [7]. Coleopterans acquire most gut microorganisms from their diet. Several of these microorganisms encode in their genome adaptations that allow them to survive in the digestive tract of their insect host. Genomic plasticity and the ability to synthetize a wide variety of molecules will secure their persistence in the gut ecosystem and allow these microbes to establish permanent symbiotic interactions with their host [7,44].

Degrading recalcitrant compounds such as cellulose and lignin is a complex and energy costly process. Xylophagous organisms survive feeding on this material while having limited access to essential nutrients including nitrogen [45]. Such conditions enhance the establishment of interactions with beneficial microorganisms that facilitate the conversion and access to some of the nutrients otherwise inaccessible. Symbiotic microorganisms are able to digest the lignocellulose material in the gut lumen while carrying out secondary processes of the carbon cycle such as methanogenesis and acetogenesis [46].

Passalid beetles can be found throughout Costa Rica and recent studies have linked their microbiota to enzymatic degradation of xylose, cellulose and lignin. Urbina and colleagues describe how Guatemalan Passalids are colonized by several Ascomycete xylose-fermenting yeast such as *Scheffersomyces shehatae* [38]. On the other hand, Costa Rican beetles harbor two Ascomycete fungus species (*Anthostomella sp* and *Hypocrea sp*), two *Streptomyces sp* and one *Bacillus sp* with high capacity to break down cell wall polymers [47]. However, these studies have been limited mostly to fungi and aerobic bacteria, despite beetle guts being mostly anaerobic and microaerophilic environments. Passalid guts possess reducing atmospheres with redox potentials below -50mV, particularly in the areas where fermentative processes occur [48]. Other studies have documented oxygen gradients throughout different areas of the digestive tract, allowing for aerobic events to happen (such as lignin degradation) [49]. Nonetheless, this occurs in the outer regions near the gut epithelium. At inner sections, where most of the plant material is located, oxygen levels drop quickly and anaerobic nucleus are created [50].

1.6 Passalid beetle microbiota research in Costa Rica

Passalid beetles are a current topic of research in Costa Rica. The project "Discovery of potential therapeutic agents and bioenergetic materials in natural products from the Costa Rican biodiversity" focused on the extraction of natural products, discovery of novel antibiotics and cellulose degrading enzymes from Costa Rica Biodiversity. It consisted on a collaboration between the National Institute of Biodiversity, the University of Costa Rica (UCR), the University of Michigan, the University of Harvard and the Joint Genome Institute. Xylophagous beetles were among the study subjects of the project. Researchers analyzed the microbial diversity of a wide array of decaying wood-associated coleopterans looking for cellulose degrading enzymes. After an initial survey of the cultured microbiota of a variety of different species, beetles from the Passalidae family were chosen as the most promising organisms to search for novel carbohydrate hydrolases. [47]

To establish the composition of the gut communities in passalid beetles,16S rRNA gene libraries were constructed followed by function analysis with shotgun metagenomics [47,51]. The experimental design for this phase of the project was the following: from five different logs, individual samples of adult, larvae and associated substrate (gallery) were collected and brought back to the laboratory to extract DNA. 16S rRNA gene libraries were sequenced by 454 pyrosequencing. After evaluating four different methodologies for DNA extraction, 15 samples (5 for each sample type) were selected for diversity analysis. Because of the rich and novel diversity of species (both archaea and bacteria) observed in the 16S RNA amplicon libraries from larvae, it was decided to sequence full metagenomes from all 5 samples of this stage and a pool of two adults and two substrates (logs 3 and 4) for a total of 7 metagenomes. All metagenomic sequences are stored in the IMG platform (https://img.jgi.doe.gov/), where they remain public and available for further use.

This previous work has four main findings: 1) Larvae, adults and gallery harbor different microbial communities, 2) Passalid larvae are enriched in OTUs from the phylum Firmicutes such as Ruminococcaceae species, 3) Passalid larvae metagenomes show a resemblance with cow rumen at a functional level (protein coding genes) and 4) Passalid gut metagenomes code for enzymes involved in cellulose degradation, methane production and nitrogen fixation. As part of his Thesis dissertation, Vargas constructed a library of 101 Metagenome Assembled Genome (MAGs) from all three sample types, and the majority of these MAGs were identified as Firmicutes [51]. However, the main limitation of the previous work is that all samples were collected at one site and from one beetle family. Thus, these results could be attributed to a site effect. Besides I wanted to compare the Passalid microbiota to other xylophagous beetles to evaluate the effect of pasalid unique ecology regardless of their diet.

The similarity at the protein level between the Passalid larval metagenome and cow rumen detected in the previous research suggests that the metabolic strategies and the taxonomic distribution of both bacterial consortiums are analogous. This idea is supported by taxonomic abundances. Passalid gut microbial communities are quite similar to cow rumen, with Firmicutes and Bacteroidetes being the most abundant phyla [52,53]. In rumen, the majority of the cellulolytic activity is carried out by anaerobic organisms from the Firmicutes and Fibrobacteres phyla [52,54,55]. Passalid metagenomes had low numbers of sequences assigned to genus Fibrobacter but high abundance of certain Firmicutes. Therefore, Firmicutes species were to be considered appropriate candidates to look for cellulose degradation pathways. In this work I decided to focus on the most abundant microbial family of Firmicutes in Passalid beetles, Ruminococcaceae.

1.7 The Ruminococcaceae family: cellulose-degrading bacteria in anaerobic gut environments.

Ruminococcaceae is a family of strict anaerobic bacteria with high morphological diversity [56] that is commonly associated with the ability to break down plant material in the digestive system of mammals [57]. The phylogeny of the Ruminococcaceae family was subjected to debate, however, based on 16S rRNA gene sequences, over 12 different genera have been described [56]. Several bacteria from the genus *Ruminococcus* are already characterized in the literature as key participants in cellulose degradation processes, both in rumen and in other environments with recalcitrant carbon sources including compost [44,52,58].

Most species in this family are described as symbionts of the intestinal microbiota of a variety of hosts, particularly herbivorous mammals. However, several species of *Ruminococcus* sp. correspond to free living organisms [59]. Regarding cow rumen, *Ruminococcus albus* and *R. flavefaciens* are two of the main species described from this type of samples. Both species are involved in decomposing processes of crystalline cellulose and the production of SCFA that will be absorbed by the host for nutrition [27,60].

In this work, I present a two-part study where first I perform a survey of the microbial community composition of Passalid beetles from different locations and compare it to different families of xylophagous beetles, Cerambycidae and Scarabaeidae. Once I confirmed that Ruminococcaceae OTUs were dominant in Passalids and other xylophagous beetles from different ecosystems, I revisited the metagenomes and MAG libraries constructed by Vargas [51] to analyze the cellulose degrading capacity of the Ruminococcaceae species present in the Passalid gut beetles. Metabolic pathways associated with cellulose degradation in Passalid beetles are still poorly comprehended. Understanding the genomic structure of this Ruminococcaceae MAGs will provide insights on the mechanisms employed to degrade this recalcitrant polymer in the Passalid beetle gut.

2. Justification

Nutrient uptake is essential for living organisms. Insects are present in almost every environment on the planet, and several species rely on symbiotic relationships to thrive. Xylophagous insects are no exception, in most cases they depend on different species of microorganisms to degrade the woody material they consume. At the same time, they take part in the turnover of carbon resources, a highly relevant process in tropical ecosystems such as those in Costa Rica. Passalid beetles excel performing this role because of their capacity to complete their whole life cycle feeding exclusively on wood from the fallen logs they inhabit. Besides, the other two beetle families included in the study are also able to degrade recalcitrant carbon sources, particularly during their larvae stage.

Most studies involving Passalid microbiota are based on the analysis of the adult microbiota, and are focused on describing the capacity to degrade fiber material by symbiotic fungi colonizing the gut [38,50]. On the other hand, few studies have focused on larvae or bacterial microbiota and even fewer in the portion of this microbiota that is anaerobic, despite the fact that sequences from anaerobe groups dominate Passalid gut ecosystems [49–51].

Our research group previously sequenced 7 metagenomes from Passalid beetles samples and generated almost 13 billion base pairs and 93 million genes. These metagenomes seem to hold a rich source of enzymes for efficient plant biomass degradation, including glycosyl hydrolases. There is a total of 8107 genes coding for glycosyl hydrolases detected in the previous analysis [51].For example, the larvae 4 metagenome encompassed more than

6400 putative genes coding for enzymes involved in the degradation of complex carbohydrates, including 1285 cellulase genes, 981 xylanase genes, 175 mannanase genes, 1822 amylase genes, 175 β-galactosidase genes, 1153 β-glucosidase genes and 871 β-xylosidase genes. The collection of Metagenome assembled genomes (MAGs) from previous research represents a valuable resource, as currently few strategies exist to analyze putative genomes from metagenomic data.

This project represents an example of how large and public accessible data, previously analyzed for other projects can be used together with peer-reviewed open software to approach a new problem. To understand the microbial communities in the gut of these insects will provide further information about how recalcitrant carbon is processed in different herbivorous insects. Further, it allows the identification of the main microorganisms involved in the metabolism of such substrates. Therefore, data presented in this dissertation will provide important information for the discovery of novel enzymes and pathways involved in cellulose degradation that could have biotechnological implications in the near future.

3. Hypothesis

Microorganisms of the Ruminococcaceae family are important components of the microbiome of xylophagous beetles, regardless of the host's native ecosystem, and their genomes encode enzymes capable of degrading cellulose allowing nutrient uptake by their host.

4. Main aim

To analyze the microbiome associated with different species of xylophagous beetles and the metagenome of organisms from the Passalidae family, to elucidate the contribution of the most abundant microbes involved in cellulose digestion.

4.1 Specific aims

i. To compare the diversity and structure of microbial communities associated with the gut of three different families of xylophagous beetles from three different ecosystems to identify dominant groups.

ii. To identify putative cellulose degrading enzymes in metagenomic sequences from members of the Ruminococcaceae family to characterize their contribution to the overall cellulose degradation by the gut microbiome.

iii. To determine the phylogenetic relationships with reference organisms of metagenome assembled genomes (MAGs) from the Ruminococcaceae family obtained from the digestive tract of Passalid beetles.

iv. To perform an in-depth analysis of the metabolic capacity of a subset of MAGs with cellulolytic potential to elucidate their possible role in the ecophysiology of Passalid Beetles.

5. Material and Methods

5.1 Samples for shotgun metagenomics and 16S rRNA gene libraries

Samples used for this study were collected between the years 2010-2013 as part of the project “Discovery of potential therapeutic agents and bioenergetic materials in natural products from the Costa Rican biodiversity”. Beetle or substrates used in the study were divided in two groups. The first one comprised every sample used for the 16S rRNA amplicon libraries. For this collection, specimens were gathered by direct encounter from the interior of decomposing fallen trees. Every insect was transported in plastic containers surrounded by gallery material and once they arrived to the lab they were dissected for DNA extraction.

As shown in Figure 5 samples were collected from the following National Parks:

1) Braulio Carrillo National Park at Central Volcanic Mountain Range Conservation Area (ACCVC): Braulio Carrillo National Park is located in Heredia and San José Province, in central Costa Rica. Both, the large size of Braulio Carrillo National Park, and its varied altitude of 3,000 meters between highest and lowest points, make it home to several distinct ecoregion zones. Ranging from high-altitude cloud forests to lowland tropical rainforest, it has one of the highest levels of biodiversity in Costa Rica. More than 90% of the park is covered in primary forest [51]

2) Corcovado National Park at OSA Conservation Area (ACOSA): The park conserves the largest primary forest on the American Pacific coastline and one of the few remaining sizable areas of lowland tropical forests in the world. Waters of the park are calm and rich in biodiversity.

3) Isla del Coco National Park at Cocos Island Conservation Area (ACMIC): It is located in the Pacific Ocean, approximately 550 km from the Pacific shore of Costa Rica. The landscape is mountainous and irregular and the summit is Cerro Iglesias at 575.5 m. Cocos Island is home to dense tropical moist forests. The island was never linked to a continent, hence the flora and fauna arrived via long distance dispersal from the Americas.

Every sampling procedure was conducted with the appropriate permits from the National Commission for Biodiversity (CONAGEBIO) by resolutions R-CM-INBio-162-2013-OT and R-CM-INBio-152-2012-OTACMIC. Cerambycidae, Scarabaeidae and Passalidae samples were collected from all three sites depending on what was found in each location. Adult, Larvae and gallery samples were collected from each beetle family included in the study. Substrate samples from each decomposing log were taken every time an insect was collected. Samples from eggs or pupae were not collected because it is not possible to identify the digestive system in these developmental stages. The taxonomic identification of each beetle was made by experts at INBio. Dissection and DNA extraction procedures were previously tested and determined to yield the best results by Vargas, 2019 [51]. Beetles were dissected in sterile conditions, using scissors and tweezers to extract and collect their gut content in phosphate buffer vials to avoid cell damage. Intestinal tissue samples were placed in 1.5ml tubes and homogenized with vortex for 10s and then sonicated for another 30 seconds. Homogenized gut contents were then placed in phosphate buffer for storage and posterior DNA extraction. Table 1 shows a total of 84 samples that were included in the downstream analysis.

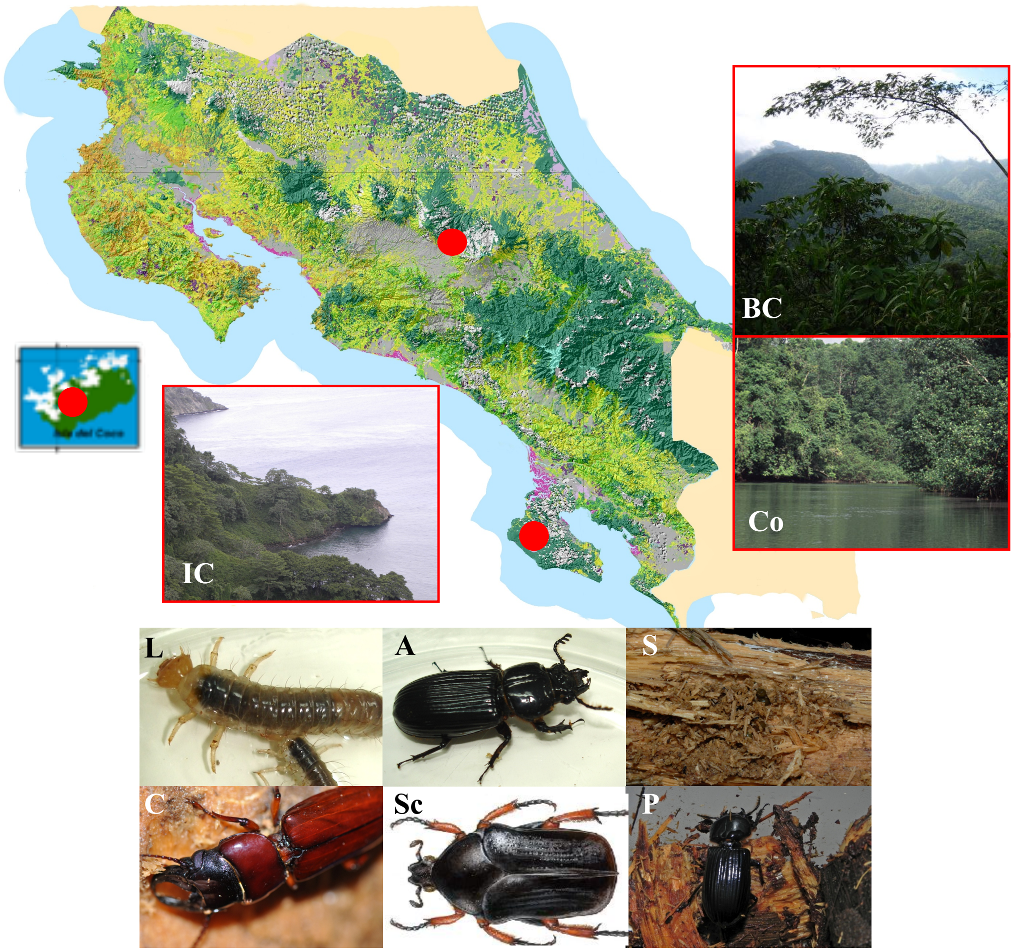


Figure 5. Sample collection scheme for 16S amplicon libraries. Insects were collected from: 1) Three sites: (IC) Isla del Coco National Park, (BC) Braulio Carrillo National Park, and (Co) Corcovado National Park. 2). Three sample types: (L) larvae, (A) adult, (S) gallery substrate material. 3) Three beetle families: (C) Cerambycidae, (Sc) Scarabaeidae, and (P) Passalidae.

23

Table 1. Metadata associated to DNA samples used in the construction of 16S rRNA genes clone libraries.

#Sample ID	Collection Date	Family	Development stage	Conservation Area	Location	Standard Longitude	Standard latitude
96343	14/3/2012	Scarabaeidae	Adult	ACCVC	Parque_Nacional Braulio_Carrillo	-83,956,583	10,148,556
96344	14/3/2012	Scarabaeidae	Adult	ACCVC	Parque_Nacional Braulio_Carrillo	-83,956,583	10,148,556
96345	14/3/2012	Scarabaeidae	Adult	ACCVC	Parque_Nacional Braulio_Carrillo	-83,953,833	10,159,722
96346	14/3/2012	Scarabaeidae	Adult	ACCVC	Parque_Nacional Braulio_Carrillo	-83,968,611	10,156,361
96347	14/3/2012	Scarabaeidae	Adult	ACCVC	Parque_Nacional Braulio_Carrillo	-83,968,083	101,265
96348	14/3/2012	Scarabaeidae	Adult	ACCVC	Parque_Nacional Braulio_Carrillo	-83,968,083	101,265
96349	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,394,361	8,532,917
96350	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,394,361	8,532,917
96351	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,394,361	8,532,917
96352	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,394,361	8,532,917
96353	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,393,833	8,531,278
96354	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,393,833	8,531,278
96355	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,393,833	8,531,278
96356	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,393,833	8,531,278
96357	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,397,472	8,531,083
96358	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,397,472	8,531,083
96359	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,397,472	8,531,083
96360	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,397,472	8,531,083
96361	29/6/2012	Passalidae	Substrate	ACOSA	Est._El_Tigre	-83,397,472	8,531,083
96362	29/6/2012	Passalidae	Adult	ACOSA	Est._El_Tigre	-83,397,472	8,531,083
96363	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,397,194	8,531,194

96364	29/6/2012	Scarabaeidae	Larvae	ACOSA	Est._El_Tigre	-83,397,194	8,531,194
96365	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,397,194	8,531,194
96366	29/6/2012	Scarabaeidae	Substrate	ACOSA	Est._El_Tigre	-83,397,194	8,531,194
96368	29/6/2012	Passalidae	Substrate	ACOSA	Est._El_Tigre	-83,397,194	8,531,194
96369	29/6/2012	Passalidae	Larvae	ACOSA	Est._El_Tigre	-83,394,694	8,532,583
96370	29/6/2012	Passalidae	Adult	ACOSA	Est._El_Tigre	-83,394,694	8,532,583
96371	30/6/2012	Passalidae	Larvae	ACOSA	Est._Los_Llanos	-83,662,806	8,641,611
96372	30/6/2012	Passalidae	Adult	ACOSA	Est._Los_Llanos	-83,662,806	8,641,611
96373	26/8/2012	Passalidae	Adult	ACCVC	Est._Quebrada_Gonzalez	-83,937,306	10,158,222
96374	26/8/2012	Passalidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,937,306	10,158,222
96375	26/8/2012	Passalidae	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,937,306	10,158,222
96376	26/8/2012	Scarabaeidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,937,306	10,158,222
96377	26/8/2012	Scarabaeidae	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,937,306	10,158,222
96378	25/8/2012	Passalidae	Adult	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96379	25/8/2012	Passalidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96380	25/8/2012	Passalidae	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96381	25/8/2012	Scarabaeidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96382	25/8/2012	Scarabaeidae	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96383	25/8/2012	Callirhypidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96384	25/8/2012	Callirhypidae	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,936,861	10,160,694
96385	25/8/2012	Passalidae	Adult	ACCVC	Est._Quebrada_Gonzalez	-83,937,222	10,160,722
96386	25/8/2012	Passalidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,937,222	10,160,722
96387	25/8/2012	Passalidae	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,937,222	10,160,722
96388	25/8/2012	Scarabaeidae	Larvae	ACCVC	Est._Quebrada_Gonzalez	-83,937,222	10,160,722
96389	25/8/2012	NA	Substrate	ACCVC	Est._Quebrada_Gonzalez	-83,937,222	10,160,722

96390	25/8/2012	Cerambycidae	Larvae	ACCVC	Est._Quebrada_G onzalez	-83,937,222	10,160,722
96391	25/8/2012	Cerambycidae	Substrate	ACCVC	Est._Quebrada_G onzalez	-83,937,222	10,160,722
96392	27/1/2013	Passalidae	Adult	ACOSA	Est._Isla_del_Can o	-83,887,028	8,705,889
96393	27/1/2013	Passalidae	Larvae	ACOSA	Est._Isla_del_Can o	-83,887,028	8,705,889
96398	28/1/2013	Passalidae	Larvae	ACOSA	Est.Los_Planes	-83,662,278	8,640,139
96399	28/1/2013	Passalidae	Substrate	ACOSA	Est.Los_Planes	-83,662,278	8,640,139
96400	28/1/2013	Scarabaeidae	Larvae	ACOSA	Est.Los_Planes	-83,662,278	8,640,139
96401	28/1/2013	Scarabaeidae	Substrate	ACOSA	Est.Los_Planes	-83,662,278	8,640,139
96402	28/1/2013	Cerambycidae	Larvae	ACOSA	Est.Los_Planes	-83,662,278	8,640,139
96403	28/1/2013	Cerambycidae	Substrate	ACOSA	Est.Los_Planes	-83,662,278	8,640,139
96404	29/1/2013	Passalidae	Adult	ACOSA	Est.Los_Planes	-83,662,806	8,640,694
96405	29/1/2013	Passalidae	Larvae	ACOSA	Est.Los_Planes	-83,662,806	8,640,694
96406	29/1/2013	Passalidae	Substrate	ACOSA	Est.Los_Planes	-83,662,806	8,640,694
96408	29/1/2013	Curculionidae	Substrate	ACOSA	Est.Los_Planes	-83,662,806	8,640,694
99904	4/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99905	4/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99906	4/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99907	4/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99908	4/6/2013	Passalidae	Adult	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99909	4/6/2013	Passalidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99910	4/6/2013	Passalidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99911	4/6/2013	Cerambycidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99912	4/6/2013	Cerambycidae	Adult	ACMIC	Parque_Nacional Isla_del_Coco	-87,054,472	5,534,361
99913	5/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556
99914	5/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556

99915	5/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556
99916	5/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556
99917	5/6/2013	Cerambycidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556
99918	5/6/2013	Passalidae	Adult	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556
99919	5/6/2013	Passalidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,575	5,532,556
99920	8/6/2013	Passalidae	Adult	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,425	5,544,333
99921	8/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,425	5,544,333
99922	8/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,425	5,544,333
99923	8/6/2013	Cerambycidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-8,705,425	5,544,333
99924	11/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,042,917	5,547,861
99925	11/6/2013	Cerambycidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,042,917	5,547,861
99926	11/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,042,917	5,547,861
99927	11/6/2013	Passalidae	Substrate	ACMIC	Parque_Nacional Isla_del_Coco	-87,042,917	5,547,861
99928	11/6/2013	Cerambycidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-87,042,917	5,547,861
99929	11/6/2013	Cerambycidae	Larvae	ACMIC	Parque_Nacional Isla_del_Coco	-87,042,917	5,547,861

5.2 Samples for shotgun metagenomics

After the confirmation of Ruminococcaceae being one of the most abundant families in the gut of Passalids, the available metagenomes were searched for evidence of cellulose degrading enzymes. For the second part of the study, seven publicly available metagenomes were downloaded (Table 2). Samples were collected during the year 2010 and correspond to 5 larvae samples, one pool of two adult guts and one pool of two substrates from family groups 3 and 4, and first analyzed by Vargas, 2019 [51]. Libraries for shotgun metagenomics were processed according to JGI protocols and sequenced in the in the Illumina HiSeq 2000 ® generating 150pb fragments. Samples statistics are summarized in Appendix 2.

Table 2. Metadata associated to DNA samples used in the construction of libraries for "whole shotgun metagenomics" sequences downloaded from IMG.

#IMG_taxon_id	Collection Date	Family	Development stage	Conservation Area	Location	Standard Longitude	Standard latitude
3300000097 (Pool Adults)	22/11/2010	Passalidae *Veturius sinuatocollis*	Adult	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151
3300000114 (Larvae 3)	22/11/2010	Passalidae *Veturius sinuatocollis*	Larvae	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151
3300000103 (Larvae 5)	22/11/2010	Passalidae *Veturius sinuatocollis*	Larvae	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151
2225789003 (Larvae 2)	22/11/2010	Passalidae *Veturius sinuatocollis*	Larvae	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151
2225789004 (Larvae 4)	22/11/2010	Passalidae *Veturius sinuatocollis*	Larvae	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151
3300000062 (Larvae 1)	22/11/2010	Passalidae *Veturius sinuatocollis*	Larvae	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151
3300000036 (Pool Gallery)	22/11/2010	Passalidae *Veturius sinuatocollis*	Substrate	ACCVC	Est._Quebrada_ Gonzalez	-84.007673	10.207151

5.3 DNA extraction for 16S rRNA gene libraries and shotgun metagenomic sequencing.

The protocol for DNA extraction employed in this work is described in Vargas, 2019. Samples were extracted with the PowerSoil® DNA Isolation Kit by Mobio. Approximately 200 µl of intestinal contents were placed in a 1.5 ml tube and treated with proteinase K in lysis buffer before continuing with the standard procedure suggested by the manufacturer. All samples were stored at -20 °C until further processing.

Vials with DNA were sent to the Joint Genome Institute (JGI) in the United States. JGI provided the sequencing service as part of their collaboration with the ICBG project. For the 16S rRNA libraries an initial PCR was made to amplify the V4 region of the gene using universal primers for Bacteria and Archaea domains 515F (5' GTGCCAGCMGCCGCGGTAA 3') and 806R (5' GGACTACHVGGGTWTCTAAT 3') [61] which are equal to those employed by the Earth Microbiome Project (http://www.earthmicrobiome.org/emp-standard-protocols/16s/) and processed in the Illumina MySeq ® generating fragments of 250pb. Samples statistics are summarized in Appendix 1.

5.4 Culture independent community diversity and structure analysis

Software package QIIME ("Quantitative Insights Into Microbial Ecology") version 1.9.1 [62] was used for the initial demultiplexing, quality filtering and taxonomic assignment of all samples. Initial quality filtering includes sequences of at least 150bp, and a minimum phred score of 25. Resulting sequences were grouped into Operational Taxonomic Units or OTUs using the algorithm UCLUST [63], the curated GreenGenes database [64] and an identity of at least 97% between sequences. After clustering, tables with the relative abundance of each OTU and their corresponding assigned taxonomy were generated.

Samples were randomly subsampled based on the one with the least number of sequences after the quality filtering step. Once all samples were rarefied to 41032 sequences, alpha diversity, beta diversity, group significance and core microbiome analyses were performed. Moreover, analysis of similarity (ANOSIM) metrics were performed on the dissimilarity matrices resulting from the beta diversity analysis. NMDS plots using Jaccard distance matrix and alpha diversity Chao1 and Shannon metrics were generated using Phyloseq [65] and vegan [66] R libraries. With the purpose of identifying OTUs that can discriminate different sample groups, a linear discriminant analysis (LDA) was conducted using LefSe software [67]. LefSe emphasizes the relevance of the effect each taxon has into a particular group using the Kruskal-Wallis, Wilcoxon and LDA statistics. Higher values of LDA correlate to higher abundance of certain taxons in a particular sample type in the study.

5.5 Carbohydrate breakdown potential of Ruminococcaceae organisms in the beetle gut metagenomes.

Assembly and annotation of the metagenomes was completed with the software and algorithms loaded into the IMG [68] platform from JGI (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi) . Subsequently, the "gene search" tool incorporated in the IMG website was used to extract the sequences from all glycosyl hydrolase (GH) genes of the selected Passalid beetle metagenomes. The resulting sequences were blasted [69] against the nucleotide database and used as an input for taxonomy assignment to each sequence with the software MEGAN [70].

Additionally, the pfam tool [71] was employed to search exclusively cellulases and cellulose binding modules (CBM) in the metagenomes. For further analyses, 4 metagenomes were selected: 2 metagenomes from individual larvae collected in decomposing logs 3 and 4, as well as the adult pool and the gallery pool of those logs. Sequences were run through

MEGAN and the ones classified as order Clostridiales and family Ruminococcaceae were extracted for the downstream analysis. Pfams for cellulases and CBM were compared against the metagenome samples including 1) all groups, 2) only Clostridiales and 3) only Ruminococcaceae. Output tables were employed to generate heatmaps with the library ggplot2 in R[72].

5.6 Analysis of Clostridiales MAGs retrieved from Passalid beetle gut metagenomic data.

Raw sequence data was downloaded from the JGI database belonging to the four metagenomes used in this study. The workflow to generate metagenome assembly genomes (MAGs) employs an informatic package named Anvi'o: Analysis and Visualization of omics data [73]. Anvi'o allows the clustering of contigs generated in the assembly and the creation of "Bins" or metagenome assembled genomes (MAGs). Raw data was trimmed to remove poor quality sequences with reads shorter than 150bp and a phred score below 30 using the software sickle[74]. Quality control was followed by processing as instructed by the Anvi'o developers (http://merenlab.org/tutorials/assembly-based-metagenomics/), this includes a combined assembly of every sample with Megahit [75], and mapping the reads against the new assembly with bowtie2 [76]. Next, all the files most be formatted for Anvi'o use.

After generating each file, Anvi'o produces a contig database and a summary of the resulting Bins as described in the "Metagenomic workflow" tutorial available in the website (http://merenlab.org/2016/06/22/anvio-tutorial-v2/). Anvi'o identifies a collection of single copy core genes (sccg) (set of essential genes found once in each genome [77]), using Hidden Markov Models (HMM) and inputs that information in the contig database.

In the next step, Anvi'o creates a profile that includes the mapping data from each sample against the assembly. Afterwards, the user combines all different profiles into a merge profile where every sample is included. During this step Anvi'o usually does the binning process using the algorithm CONCOCT [77]. In this work, instead of using CONCOCT , the bin collection was obtained employing the algorithm in the software MetaBAT [78] that yielded both more complete and larger bins. The results from MetaBAT are later imported to the Anvi'o merged profile. Finally, Anvi'o includes an option where each Bin generated with MetaBAT can be refined to eliminate contaminating contigs from other organisms. When the refining and selection steps of the bins are completed, a summary is generated that includes length, N50 and number of contigs together with completion and redundancy metrics to evaluate the quality of each Bin.

5.7 Phylogenomic analysis of bins and assignment of MAGs into the Ruminococcaceae family.

After summarizing every bin generated with Anvi'o and MetaBAT, bins with over 70% completion, less than 10% redundancy and over 1MB of length were selected for downstream analysis as previously described[51] and according to standards for high quality MAGs defined in recent studies[79–82]. From the collection of MAGs, the first step was to assign a preliminary taxonomy. Every single copy core gene generated with Anvi'o was extracted and blasted against the Kbase bacteria and archaea reference genomes database[83] using the software DIAMOND [84]. Bins with over 20% of the genes assigned to the Ruminococcaceae family, 20% to the Lachnospiraceae family or at least 30% to Clostridia class were selected for the phylogenomic analysis. Reference genomes from type strains of Ruminococcaceae and Lachnospiraceae organisms along with other Clostridia groups were selected from a recent review on Ruminococcus taxonomy [59] (Table 3). Both reference genomes and selected MAGs were input in the Anvi'o tutorial for phylogenomics (http://merenlab.org/2017/06/07/phylogenomics/) to generate a maximum likelihood tree using a concatenated collection of at least 65 single copy core genes (sccg). Employing this tree, final taxonomy assignments for each MAG were confirmed. For the tree to be completed, three *Bacillus* reference genomes were selected as root.

Table 3. Reference genomes employed for phylogenomic and pangenomic analysis.

Genome Name / Sample Name	Sequencing Center	IMG Genome ID	Genome Size	Gene Count
FAMILY_RUMINOCOCCACEAE				
Ruminococcus albus DSM 20455	DOE Joint Genome Institute (JGI)	2558860976	4332295	4026
Subdoligranulum variabile DSM 15176	Washington University in St. Louis	2562617078	3237471	3450
Ruminococcus flavefaciens YL228	DOE Joint Genome Institute (JGI)	2593339237	3365971	2890
Ruminococcus bicirculans 80/3	Institute of Food Research	2585427614	2968500	2652
Ruminococcus albus 8	J. Craig Venter Institute (JCVI)	2562617137	4052160	3897
Ruminococcus callidus ATCC 27760	Washington University in St. Louis	2597490297	3070687	2913
Ruminococcus bromii YE282	DOE Joint Genome Institute (JGI)	2593339225	2539482	2571
Ruminococcus champanellensis JCM 17042	University of Tokyo	2734481917	2511227	3107
Ruminococcus flavefaciens ND2009	DOE Joint Genome Institute (JGI)	2558309005	3636694	3209
Ruminococcus albus AR67	DOE Joint Genome Institute (JGI)	2593339152	4276527	3887
Ruminococcus flavefaciens XPD3002	DOE Joint Genome Institute (JGI)	2606217757	3642793	3244
Clostridium sporosphaeroides DSM 1294	DOE Joint Genome Institute (JGI)	2519899647	3176598	3019
FAMILY LACHNOSPIRACEAE				
Ruminococcus gnavus CC55_001C	Broad Institute	2558860359	3177226	3041
OTHER CLOSTRIDIALES				
Eubacterium hallii L2-7	Wageningen University	2775506742	3515670	3336
Clostridium cellulosi DG5	Center for Biotechnology (CeBiTec), Bielefeld Univeristy	2639762740	2229578	2076

Clostridium sporogenes DSM 795	Okinawa Institute of Science and Technology	2636415665	4142990	3890
Clostridium botulinum CDC_53174	Los Alamos National Laboratory	2718218005	3867627	3599
Clostridium carboxidivorans P7	Shanghai Institutes for Biological Sciences (SIBS) of Chinese Academy of Sciences (CAS)	2654588030	5752782	5316
ROOT_Bacillus				
Bacillus subtilis subtilis AG1839	Massachusetts Institute of Technology	2585427616	4193640	4347
Bacillus cereus ATCC 10876	Los Alamos National Laboratory	2609460101	5993683	6108
Bacillus cereus MIT0214	Massachusetts Institute of Technology	2627854006	5594243	5788

5.8 Pangenomic analysis of protein clusters common to the Ruminococcaceae family and associated to cellulose breakdown in Passalid beetles

Anvi'o has a tool that performs a pangenomic comparative analysis of all proteins coded in each MAG included in the analysis (http://merenlab.org/2016/11/08/pangenomics-v2/). Beside the ones generated during the merge profile step, Anvi'o admits the inclusion of related reference genomes. It starts by generating a collection including all MAGs and genomes to evaluate. Each genome or MAG was annotated using the COG database to retrieve the functional annotation of every gene. Afterwards a pan-database is generated. Employing each database, the program compares each protein's similarity using BLAST. Next, clusters of proteins are build based on how similar they are to each other using the algorithm MLC and a inflation value of 1 [85]. Weak clusters were eliminated through Maxbit heuristic observations with ITEP [86] to ensure only confident clusters were kept for further analysis. Finally, the results can be managed in the same interactive platform used to refine bins, but in this case was used to select protein clusters associated with a particular phylogenetic group inside the Ruminococcaceae family. Tables with these results including every gene call and their putative COG functions were generated for further discussion.

Besides using Anvi'o, every MAG assigned to the Ruminococcaceae family was annotated with PRODIGAL [87] and the resulting amino acid sequences were used as an input in the dbCan [88] online tool to identify every hit of carbohydrate active enzymes from the CAZy database [23] contained in the genomes. A table with every genome hits for CAZy enzyme was generated as a result.

5.9 Analysis of main metabolic pathways and carbohydrate degrading capacity in selected MAGs.

Bin.519, Bin.503, Bin.86 and Bin.174 were selected for further analysis of their metabolic pathways to evaluate their overall metabolic potential. They were chosen because putative cellulases were identified in their genomes from the pangenomic study. Annotation of these MAGs was completed with the online tool Rast (http://rast.nmpdr.org/rast.cgi) [89–91] and Kegg annotation tools [92]. Annotated genomes were revised with Artemis and DNA Plotter [93] to identify possible genome clusters for cellulose degradation.

6 Results.

6.1 Community structure of xylophagous beetle microbiomes

6.1.1 Analysis of all 84 samples included in the survey.

Three different variables were evaluated to characterize the microbiomes of the Passalid and other xylophagous beetles: 1) Geographic location, 2) Beetle family and 3) Sample type. Altogether, samples generated a total of 6293997 reads after sequencing and quality filtering. A total of 8375 different OTUs were identified in this study. Samples had an average of 73186 reads with a minimum of 41032 and a maximum of 97505. A summary of all samples is shown in Annex 1.

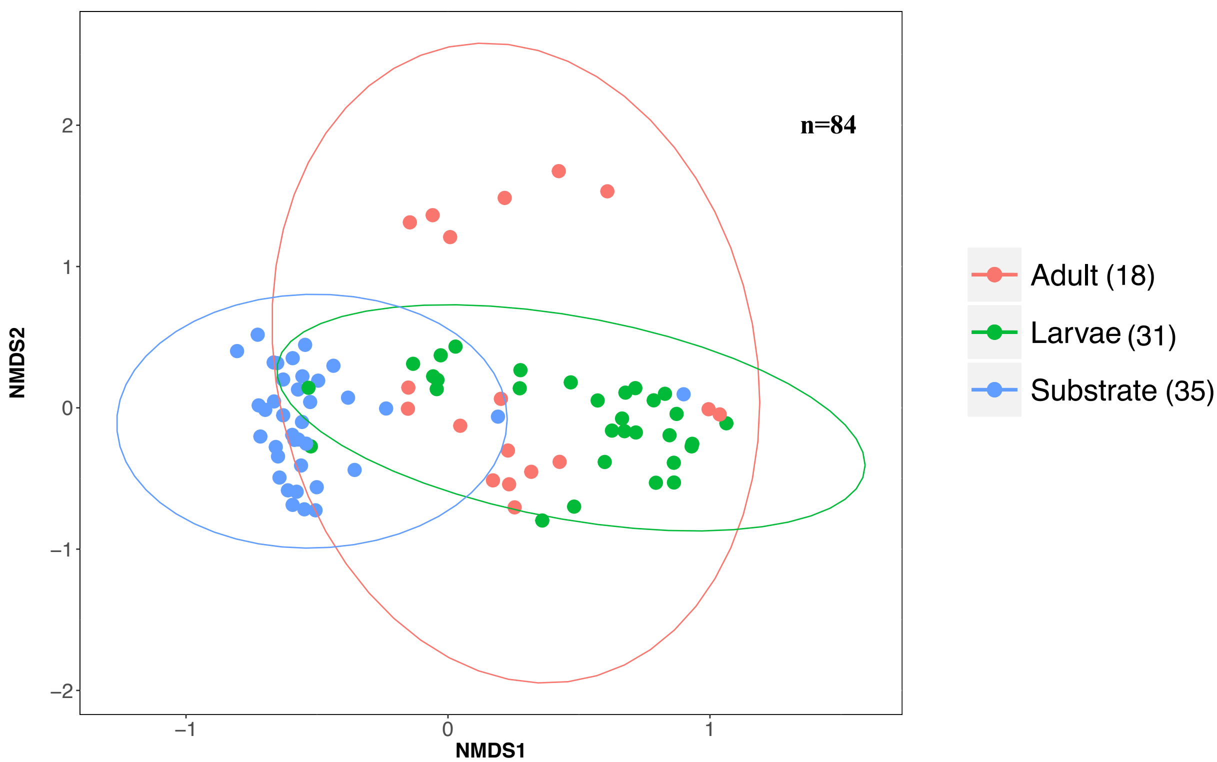


Figure 6. NMDS plot for all samples analyzed in the study. Each dot on the plot corresponds to a different microbial community. Samples are colored according to type: larvae (green), adults (red) and substrate (blue). ANOSIM and PERMANOVA analysis showed that host sample type, host family and host geographic location are significant components of the variation of microbial

communities. A normal distribution for each group is shown by a continuous line. Numbers in parenthesis correspond to the sample size within each group. Stress value of NMDS was 0,11.

Table 4. Beta diversity statistics for all samples grouped by site, beetle family and sample type

	Site	**Beetle family**	**Sample type**
Sample size	84	84	84
Number of groups	3	3	3
Test statistic	0,302	0,339	0,417
p-value	**<0,010**	**<0,010**	**<0,010**
Number of permutations	99	99	99

First, I analyzed every sample together to understand the similarities amongst microbiomes in the three xylophagous beetle families under study. As shown in the NMDS plot in Fig. 6, the major trait that groups samples are the host sample type. For example, despite being from different families of beetles, all the larval gut microbiomes grouped close to each other. The substrate samples are also clustered near each other and apart from the rest of the samples, suggesting that the beetle families had little effect over their surrounding microbiota. On the other hand, adults from different families seem to have more unequal communities from one another, even some clustering closer to larvae samples. Even so, host sample, host family and host geographic location also showed statistically significant effects over the microbial communities with the ANOSIM test (Table 4).

Alpha diversity metrics, Chao1 and Shannon, were used to evaluate how rich and diverse are gut communities according to sample type. Adult and larval guts seem to be equally rich, while the substrate showed higher values (Figure 7) according to the Chao1 index. Samples were not different according to beetle family. Regarding the sample source, only samples from ACCVC seem to be richer than samples from ACOSA. ACMIC samples did not differ from any of the others according to statistical analyses shown in table 5. All three sample types have statistical differences in diversity (table 5), as larvae samples are more diverse that adults, but less diverse than substrate (Figure 7). Consistent with the richness index, ACCVC microbiomes seem to be more diverse than ACOSA but statistically similar to ACMIC.

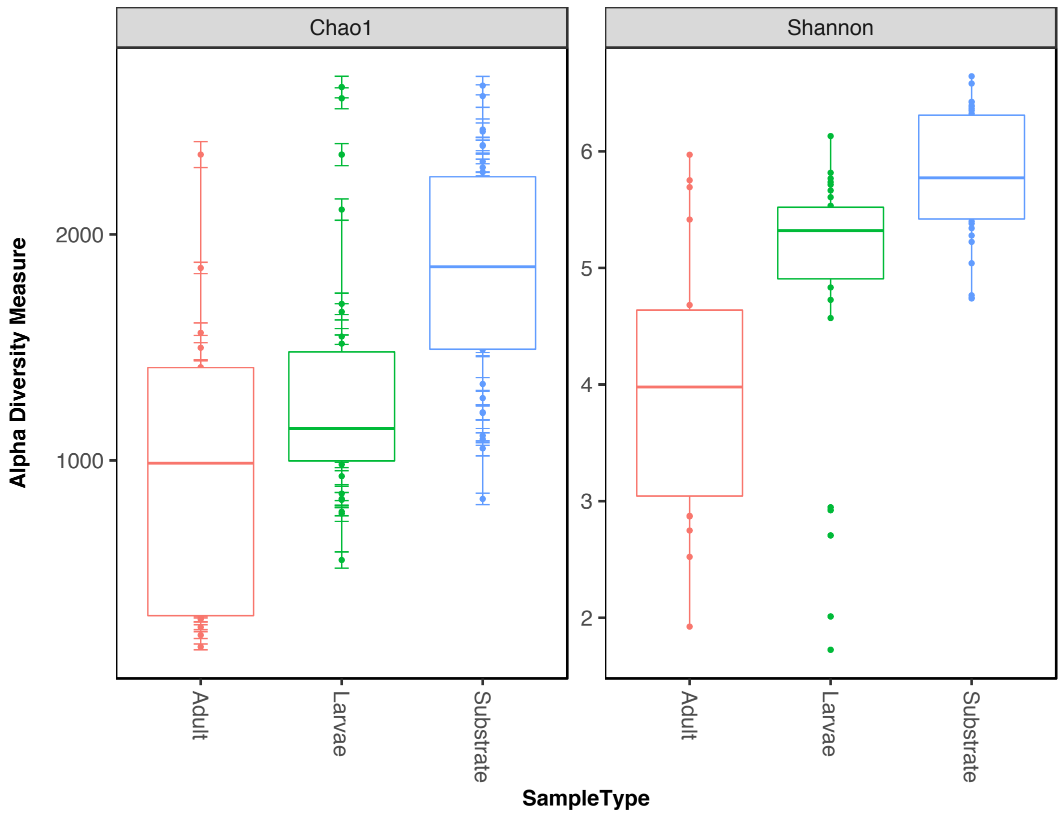


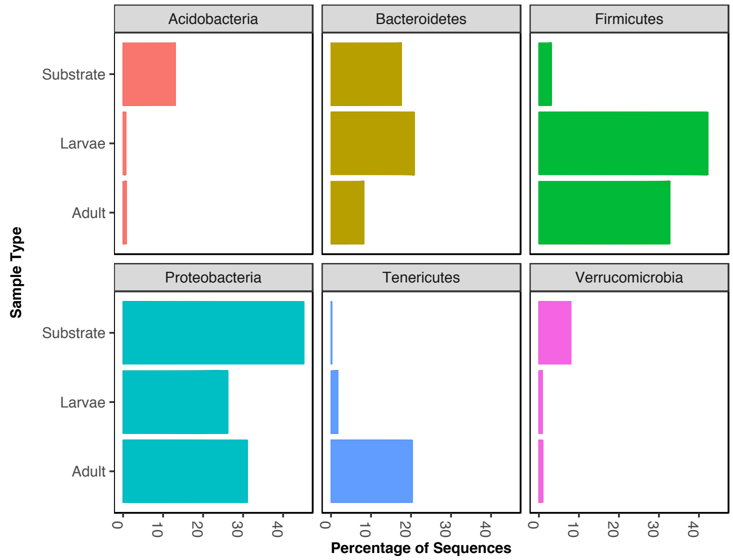
Figure 7. Boxplots for Chao1 and Shannon alpha diversity index for all samples analyzed in the study (n=84).

Table 5. Statistics for alpha diversity index for all samples grouped by site, beetle family and sample type.

Chao1 All Samples Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	-2.628	0.057
ACMIC	ACOSA	-2.143	0.108
ACCVC	ACOSA	-4.193	**0.003**
Chao1 All Samples Beetle Family			
Group1	Group2	t stat	p-value
Scarabaeidae	Cerambycidae	1.046	0.945
Cerambycidae	Passalidae	-2.002	0.183
Scarabaeidae	Passalidae	-0.357	1.000
Chao1 All Samples Sample Type			
Group1	Group2	t stat	p-value
Substrate	Adult	5.332	**0.003**
Substrate	Larvae	4.055	**0.003**
Adult	Larvae	-2.216	0.141
Shannon All Samples Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	-1.824	0.204
ACMIC	ACOSA	-0.715	1.000
ACCVC	ACOSA	-2.697	**0.030**
Shannon All Samples Beetle Family			
Group1	Group2	t stat	p-value
Scarabaeidae	Cerambycidae	1.368	0.528
Cerambycidae	Passalidae	-2.375	0.066
Scarabaeidae	Passalidae	-0.901	1.000
Shannon All Samples Sample Type			
Group1	Group2	t stat	p-value
Substrate	Adult	7.614	**0.003**
Substrate	Larvae	4.222	**0.003**
Adult	Larvae	-2.545	**0.045**

Taxonomic profiles for the six more abundant phyla in every sample shown in Figure 8A indicate that Firmicutes is the most abundant phylum in larvae (40%) and adults (30%). Bacteroidetes showed important abundances in both substrate (19%) and larvae (21%), but not in adults (6%). Likewise, Acidobacteria is prevalent only in the substrate (10%). On the other hand, Proteobacteria was present in every sample, but is the most abundant in substrate samples with almost 45%. Another dominant group, Tenericutes, had very low relative abundance in larvae and substrate but represented 20% of adult microbiomes. As shown in Figure 8B, Ruminococcaceae is the most abundant family in larvae with 15% follow by Lachnospiraceae (7%), both belonging to the phylum Firmicutes. Lachnospiraceae is more abundant in adult samples (9%). Enterobacteriaceae is highly abundant in larvae (8%) and adults (5%), ranking as the main representative of the Proteobacteria Phylum. The most abundant families in Bacteroidetes are Porphyromonadaceae, highly abundant in larvae (12%) and Chitinophagaceae, the most abundant in substrate (9%). Finally, the most abundant family in Adults is Pseudomonadaceae (7%).

A



B

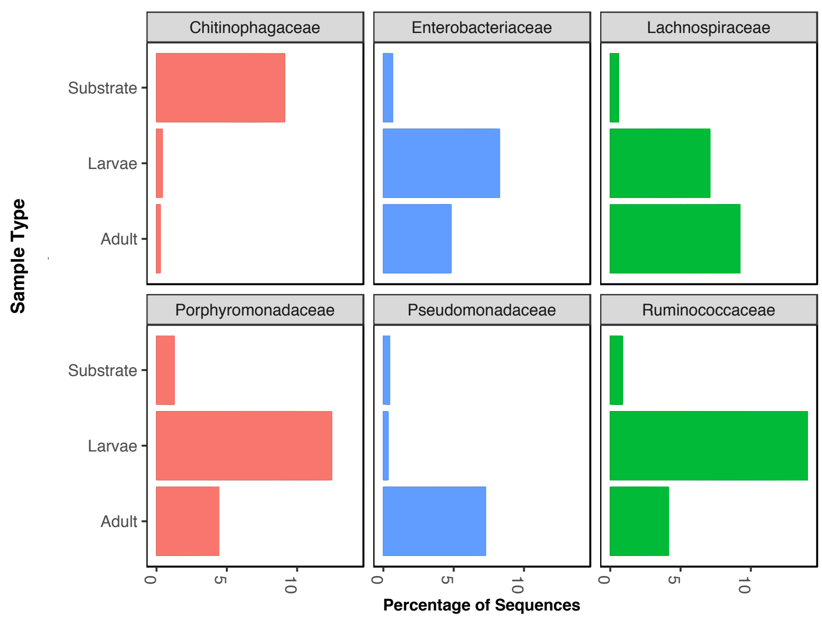


Figure 8. Relative abundance of the six most represented phyla (A) and families (B) for all samples analyzed in the study (n=84).

6.1.2 Analysis of samples associated to the Passalidae beetle family.

Previous results provided evidence that Passalid larvae host different communities than adults and substrate, confirming these findings is one of the objectives of this work. The NMDS plot in Figure 9A illustrates how the three sample types associated with the Passalidae family clustered together according to sample type, despite their geographic origin. Also, statistical analysis of the beta diversity data (Table 6) shows statistical differences in their bacterial communities both by sample type and geographic sites. Larvae samples are clustered closer together than the other two groups. Substrate and adult samples make a defined cluster with just two exceptions that fall out of the normal distribution. Adult and larvae samples seem to have similar richness and different diversity when all samples are analyzed together (Fig. 10A). Also, substrate microbiomes were the most diverse and rich overall. Finally, every geographic location tends to have the same richness and diversity according to the statistical analysis (Table 7).

Table 6. Beta diversity statistics for Passalidae samples by site and sample type.

	Site	**Sample type**
Sample size	36	36
Number of groups	3	3
Test statistic	0,172	0,590
p-value	**<0,010**	**<0,010**
Number of permutations	99	99

Table 7. Statistics for alpha diversity index for Passalidae samples by site and sample type.

Chao1 Passalidae Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	-0.493	1.000
ACMIC	ACOSA	-0.260	1.000
ACCVC	ACOSA	-0.629	1.000
Chao1 Passalidae Sample Type			
Group1	Group2	t stat	p-value
Substrate	Adult	3.875	**0.009**
Substrate	Larvae	4.974	**0.003**
Adult	Larvae	0.719	1.000
Shannon Passalidae Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	-1.267	0.663
ACMIC	ACOSA	1.0673	0.885
ACCVC	ACOSA	-0.0420	1.000
Shannon Passalidae Sample Type			
Group1	Group2	t stat	p-value
Substrate	Adult	5.501	**0.003**
Substrate	Larvae	3.646	**0.003**
Adult	Larvae	-3.526	**0.006**

The most abundant phylum in both adults and larval samples belonging to Passalid beetles was Firmicutes (Figure 11A). In the larvae, this group comprises half of the relative abundance present. Meanwhile, the most abundant groups in the substrate were Proteobacteria, followed by Bacteroidetes, Acidobacteria and Verrucomicrobia. The most striking differences between larvae and adults were given by the second most abundant group. In larvae, the Bacteroidetes are the second most abundant group (32%), while in adults it was Tenericutes (37%). The sixth more abundant families amongst samples are plotted in Fig 12A. The most abundant family in larvae was Ruminococcaceae (16%) which had a very similar abundance to Porphymonadaceae, and Lachnospiraceae was the third most abundant (12%). In contrast, adults have Lachnospiraceae as the most abundant family with 15%, followed by Ruminococcaceae (5%) and Porphymonadaceae (4%).

I employed linear discriminant analysis (LDA) of effect size (LEfSe) to determine the most significant differences between adults and larvae from Passalid samples (Figure 14). Unclassified Ruminococcaceae OTUs were found to have higher scores in larvae compared to adults. Besides, unclassified Clostridiales and Candidatus *Azobacteroides* are also more represented in larvae. On the other hand, enrichment of unclassified Mollicutes OTUs were found in adult samples when compared to larvae.

6.1.3 Analysis of samples associated to the Cerambycidae beetle family.

Table 8. Beta diversity statistics for Cerambycidae samples by site and sample type.

	Site	Sample type
Sample size	17	17
Number of groups	3	3
Test statistic	0,415	0,178
p-value	**<0,010**	**<0,050**
Number of permutations	99	99

Cerambycidae samples had only one adult specimen, therefore the beta analysis (Figure 9B) only shows two clusters, one for substrate and one for larvae. Out of all substrate samples, only one fell off the normal distribution, while larvae samples clustered within it. Table 8 shows that both geographical site and sample type are different among samples despite adults having one sample.

Richness for every sample type was statistically the same (Table 9). The statistical analysis indicates that substrate is more diverse than larvae in terms of the Shannon index, which was the only statistical difference found in alpha diversity among this group.

Table 9. Statistics for alpha diversity index for Cerambycidae samples by site and sample type.

Chao1 Cerambycidae Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	-0.768	1.000
ACMIC	ACOSA	2.206	0.165
ACCVC	ACOSA	1.242	1.000
Chao1 Cerambycidae Sample Type			
Group1	Group2	t stat	p-value
Substrate	Larvae	1.257	0.774
Shannon Cerambycidae Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	-1.098	0.750
ACMIC	ACOSA	0.949	0.978
ACCVC	ACOSA	-0.131	1.000
Shannon Cerambycidae Sample Type			
Group1	Group2	t stat	p-value
Substrate	Larvae	3.357	**0.024**

Cerambycidae samples are more similar to each other in terms of abundance as Figure 11B illustrates. In this case, Proteobacteria was the most prevalent phylum and its abundance is similar between larvae (50%), adults (41%) and substrate (43%); this is also the case for Bacteroidetes and Verrucomicrobia. Actinobacteria and Acidobacteria phyla were more abundant in substrate (9%) than adults and larvae. On the contrary, Firmicutes is more abundant in adults (23%) and larvae (21%). The family Enterobacteriaceae (Figure 12B) is highly prevalent among larvae samples (33%) when compared to adults and substrate samples; altogether with family Porphymonadaceae (10%) and Xanthomonaceae they dominate the larval gut. Chitinophagaceae family was more abundant in substrate samples while Moraxellaceae was the most abundant in adult samples. Ruminococcaceae family was present in all three samples albeit at low abundances below 5%.

6.1.4 Analysis of samples associated to the Scarabaeidae beetle family.

Beta diversity analysis for Scarabaeidae beetles illustrates how all three sample types clustered apart from each other in the NMDS plot (Figure 9C). Besides, almost every sample was inside the normal distribution. No Scarabaeidae samples were found in Cocos Island, therefore only two sites were compared, regarding alpha and beta diversity analysis. Both, sample type and geographic site show significant differences amongst samples. Adults had a lower richness compared to larvae and substrate, as shown in Figure 10C and Table 11. Furthermore, there are no differences in richness or diversity regarding the collection site between ACCVC and ACOSA. Scarabaeidae substrate samples had the highest diversity, followed by larvae and then adults, with the lowest Shannon index.

Table 10. Beta diversity statistics for Scarabaeidae samples by site and sample type

	Site	**Sample type**
Sample size	31	31
Number of groups	2	3
Test statistic	0,520	0,702
p-value	**<0,010**	**<0,010**
Number of permutations	99	99

Table 11. Statistics for alpha diversity index for Scarabaeidae samples by site and sample type.

Chao1 Scarabaeidae Site			
Group1	Group2	t stat	p-value
ACCVC	ACOSA	-5.903	**0.001**
Chao1 Scarabaeidae Sample Type			
Group1	Group2	t stat	p-value
Substrate	Adult	11.042	**0.003**
Substrate	Larvae	2.279	0.126
Adult	Larvae	-5.077	**0.003**
Shannon Scarabaeidae Site			
Group1	Group2	t stat	p-value
ACCVC	ACOSA	-3.904	**0.001**
Shannon Scarabaeidae Sample Type			
Group1	Group2	t stat	p-value
Substrate	Adult	11.001	**0.003**
Substrate	Larvae	4.151	**0.003**
Adult	Larvae	-8.177	**0.003**

For Scarabaeidae samples, Proteobacteria was the most abundant phylum overall (Figure 11C). Adults had an abundance of 60%, larvae over 20% and substrate almost 50%. Firmicutes were mostly prevalent in larvae (40%), similar to Passalids, followed by adults (23%). In this case, Bacteroidetes was similarly distributed amongst every sample type. Three different families of Proteobacteria were most abundant in adult samples when compared to larvae and substrate (Figure 12C): Pseudoalteromonadaceae (10%), Enterobacteriaceae (13%) and Pseudomonadaceae (15%). For the larval samples, the two most abundant families were Ruminococcaceae (17%) for Firmicutes and Porphyromonadaceae (13%) for Bacteroidetes. For substrate, the most abundant family was Chitinophagaceae (9 %).

6.1.5 **Analysis of xylophagous beetles larval samples.**

Larvae microbiome samples from Passalidae had very similar community compositions as they are scattered closely to each other (Figure 9D). Scarabaeidae samples are more dispersed but they also cluster apart from the other two families. Cerambycidae microbiomes had the biggest t-distribution ellipse and while four samples cluster separately, three of them were close to Passalid samples. For this analysis, site and beetle family microbiomes were statistically different. Alpha diversity analysis for all three groups (Figure 10D) showed a larger richness in about half of Scarabaeidae samples, meanwhile the other half had similar values with the other two groups; however, neither site, nor beetle family, were statistically different (Table 13). Shannon diversity index indicates higher values for both Scarabaeidae and Passalidae samples, supported by statistics (Table 13), suggesting a higher evenness and less variation compared to Cerambycidae larvae.

Table 12. Beta diversity statistics for larvae samples by site and beetle family.

	Site	**Beetle family**
Sample size	31	31
Number of groups	3	3
Test statistic	0,354	0,719
p-value	**0,010**	**0,010**
Number of permutations	99	99

Table 13. Statistics for alpha diversity index for larvae samples by site and beetle family.

Chao1 Larvae Site			
Group1	Group2	t stat	**p-value**
ACCVC	ACMIC	-2,857	0,039
ACMIC	ACOSA	-1,209	0,783
ACCVC	ACOSA	-2,590	0,054
Chao1 Larvae Beetle Family			
Group1	Group2	t stat	**p-value**
Scarabaeidae	Cerambycidae	1,831	0,219
Cerambycidae	Passalidae	-1,036	0,963
Scarabaeidae	Passalidae	1,601	0,384
Shannon Larvae Site			
Group1	Group2	t stat	**p-value**
ACCVC	ACMIC	0,443	1,000
ACMIC	ACOSA	-1,782	0,237
ACCVC	ACOSA	-1,305	0,618
Shannon Larvae Beetle Family			
Group1	Group2	t stat	**p-value**
Scarabaeidae	Cerambycidae	3,940	**0,009**
Cerambycidae	Passalidae	-4,119	**0,009**
Scarabaeidae	Passalidae	-0,988	1,000

Taxa distribution shown in Figure 11D indicates dominance of Firmicutes phylum in both Passalidae (46%) and Scarabaeidae (38%), whereas the most representative phylum in Cerambycidae was Proteobacteria. Bacteroidetes was evenly distributed amongst all three families and the Archaeal phylum Euryarchaeota was moderately more represented in Passalidae when compared with the other two groups. The taxonomy plot (Figure 12D) illustrates how Ruminococcaceae family is the most abundant family in Passalidae and Scarabaeidae. Scarabaeidae samples had a higher representation of Desulfovibrionaceae OTUs and a higher abundance of Rikenellaceae than the other two families. As stated, Cerambycidae larvae seem to be dominated by Proteobacteria, and in this case, the most abundant family present in the samples was Enterobacteriaceae.

Analysis of the core microbiome was performed for all larvae gut samples (Table 14). Depending on the beetle family, different number of core OTUs were present for each group. All larvae included in this study shared six OTUS on at least 90% of the samples, however not a single OTU was shared amongst all samples. Cerambycidae had the lower amount shared OTUs amongst the three families, with four of them in a 100% and 90% of samples, whereas Scarabaeidae had a slightly higher number with 22 OTUs in all samples and 47 OTUs shared in 90% of Sacarabaeidae gut samples. Passalidae gut samples had the greater total core OTUs with 146 in 100% of the samples and 286 in at least 90% of them. Passalid larvae had the most abundant core microbiome, dominated by phylum Firmicutes, Bacteroidetes and Euryarchaeota. Amongst samples, the Phylum Proteobacteria, dominated the core microbiome of all larvae samples.

Table 14. Number of OTUs in the core microbiome of larvae samples.

	Larvae			
	All	Passalidae	Cerambycidae	Scarabaeidae
Number of Otus in 70% core microbiome	38	528	91	150
Number of Otus in 90% core microbiome	6	286	4	47
Number of Otus in 100% core microbiome	0	146	4	22
Number of samples	31	11	7	13

6.1.6 **Analysis of adult samples associated to xylophagous beetles.**

Adult samples have different gut microbial communities as confirmed by statistical analyses (Table 15). Passalid and Scarabid adults clustered together in opposite sides of the plot displayed in Figure 9E. Scarabaeidae are both less rich and less diverse than Passalidae samples as shown in Table 16. Finally, the adult gut samples from the ACCVC area appear to be more diverse that the samples from ACMIC.

Table 15. Beta diversity statistics for adult samples by site and beetle family.

	Site	**Beetle family**
Sample size	18	18
Number of groups	3	3
Test statistic	0,266	0,866
p-value	**0,020**	**<0,010**
Number of permutations	99	99

Table 16. Statistics for alpha diversity index for adult samples by site and beetle family.

Chao1 Adult Site			
Group1	Group2	t stat	**p-value**
ACCVC	ACMIC	-2.387	0.120
ACMIC	ACOSA	0.011	1.000
ACCVC	ACOSA	-2.462	0.105
Chao1 Adult Beetle Family			
Group1	Group2	t stat	**p-value**
Scarabaeidae	Passalidae	-7.038	**0.003**
Shannon Adult Site			
Group1	Group2	t stat	**p-value**
ACCVC	ACMIC	-4.281	**0.012**
ACMIC	ACOSA	2.679	0.123
ACCVC	ACOSA	-1.332	0.612
Shannon Adult Beetle Family			
Group1	Group2	t stat	**p-value**
Scarabaeidae	Passalidae	-3.243	**0.021**

Taxonomic distribution of adult samples in Figure 11E also highlights differences between all three families. Passalidae were enriched in Firmicutes (38%) and Tenericutes (26%). However, Tenericutes was not prevalent in the other two beetle families. Proteobacteria is prevalent in both Scarabaeidae (60%) and Cerambycidae (41%) adults, unlike Passalidae. Euryarchaeota is present in all three beetle families but was found in higher proportions in Passalidae than in Scarabaeidae and Cerambycidae. The most abundant families of bacteria in Scarabaeidae were: Pseudomonadaceae and Enterobacteriaceae for phylum Proteobacteria and the Clostridiaceae family for phylum Firmicutes (Figure 12E). Passalid adults have a high abundance of Lachnospiraceae and a similar distribution of Clostridiaceae. The overall abundance of different groups of bacteria in the core microbiome for adults is similar to what was found in the larvae. As stated in Table 17, Passalidae species have the highest number of OTUs found in 70, 90 and 100%

of the samples, indicating a more stable community than Scarabaeidae. Cerambycidae adult core microbiome could not be calculated as a consequence of the low number of samples. However, the core microbiome of every adult sample seems to be composed of only one Proteobacteria OTU (Figure 13B). Meanwhile, Scarabaeidae and Passalidae have similar core microbiomes. The only difference found is that the Passalid core had more OTUs than Scarabaeidae and had a high abundance of Tenericutes OTUs.

Table 17. Number of OTUs in the core microbiome of adult samples.

	Adults		
	All	Passalidae	Scarabaeidae
Number of Otus in 70% core microbiome	9	222	52
Number of Otus in 90% core microbiome	1	53	16
Number of Otus in 100% core microbiome	0	21	16
Number of samples	18	11	6

6.1.7 **Analysis of substrate samples associated to xylophagous beetles.**

Substrate samples clustered together by beetle family, but at the same time, the NMDS analysis in Figure 9F shows a lot of overlapping between the different origins. Yet, ANOSIM statistical analysis (Table 18) indicates that geographical site and beetle family are predictors of the bacterial community composition and their differences within groups. Cerambycidae associated substrate, as shown in Figure 10F and corroborated in Table 19, were less rich in different OTUs and less diverse than both Passalidae and Scarabaeidae samples. Interestingly, Passalid and Scarabid samples have almost the same distribution regarding alpha diversity. The statistical analysis also shows that ACMIC samples had lower richness than ACOSA samples but were equally diverse.

Table 18. Statistics for alpha diversity index for adult samples by site and beetle family.

	Site	**Beetle family**
Sample size	35	35
Number of groups	3	3
Test statistic	0,496	0,230
p-value	<0,010	<0,010
Number of permutations	99	99

Substrate samples, contrary to what has been observed in the other sample types, had very similar taxonomic abundances as seen in Figure 11F. Proteobacteria is the main phylum in all three beetle families associated substrates, with more than 40% in each. Afterwards, Bacteroidetes and Acidobacteria are the two most prevalent. Firmicutes, usually predominant in gut samples of adults and larvae, had a lower abundance in the substrate. The Bacteroidetes family Chitinophagaceae is the most prevalent in the three types of substrate: Cerambycidae (11%), Passalidae (9%) and Scarabaeidae (7%). Next, the family Acidobacteriaceae is the second most abundant amongst all substrates (Figure 12F), and is more prevalent in substrate associated to Passalidae. Three different families of Proteobacteria were among the sixth most abundant families found in the substrate. Hyphomicrobiacae and Sinobacteraceae were more abundant in Passalidae, while Rhodospirillaceae was the most abundant in the Scarabaeidae substrate. Finally, family auto67_4W, a group of uncultured Verrucomicrobia species seem to be present in every sample in similar abundances.

Substrate samples share the highest number of OTUs, 18 of them shared amongst at least 90% of these samples. However, Passalidae substrate had the most OTUs on 70% of the samples with 381 OTUs, while Scarabaeidae shared 48 OTUs within 90% of samples, meanwhile Cerambycidae had 28 OTUs present in all samples. All18 OTUs shared by the substrate samples belong to the Proteobacteria phylum (Figure 13C). On the other hand, the core microbiome for each beetle family substrate had other groups present. In this regard, Cerambycidae core had a high abundance of Actinobacteria, while Acidobacteria was prevalent in Passalidae substrate.

Table 19. Statistics for alpha diversity index for adult samples by site and beetle family.

Chao1 Substrate Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	0.816	1.000
ACMIC	ACOSA	-2.960	**0.036**
ACCVC	ACOSA	-1.761	0.321
Chao1 Substrate Beetle Family			
Group1	Group2	t stat	p-value
Scarabaeidae	Cerambycidae	3.820	**0.012**
Cerambycidae	Passalidae	-3.317	**0.012**
Scarabaeidae	Passalidae	0.486	1.000
Shannon Substrate Site			
Group1	Group2	t stat	p-value
ACCVC	ACMIC	1.576	0.375
ACMIC	ACOSA	-2.199	0.075
ACCVC	ACOSA	-0.393	1.000
Shannon Substrate Beetle Family			
Group1	Group2	t stat	p-value
Scarabaeidae	Cerambycidae	3.461	**0.015**
Cerambycidae	Passalidae	-3.232	**0.012**
Scarabaeidae	Passalidae	-0.048	1.000

Table 20. Number of OTUs in core microbiome of substrate samples

	Substrate			
	All	Passalidae	Cerambycidae	Scarabaeidae
Number of Otus in 70% core microbiome	143	381	178	222
Number of Otus in 90% core microbiome	18	38	28	48
Number of Otus in 100% core microbiome	2	8	28	24
Number of samples	35	14	9	12

Having shown that Ruminococcaceae OTUs are indeed consistently associated with xylophagous beetle larvae from different families (Scarabaeidae and Passalidae) and throughout different ecosystems, I now turn my focus to study the cellulolytic potential of these microorganisms. Freely available metagenomes and the MAG collection described in previous work [51] were employed to investigate how microorganisms of the Ruminococcaceae family break down cellulose inside the gut of Passalidae beetles.

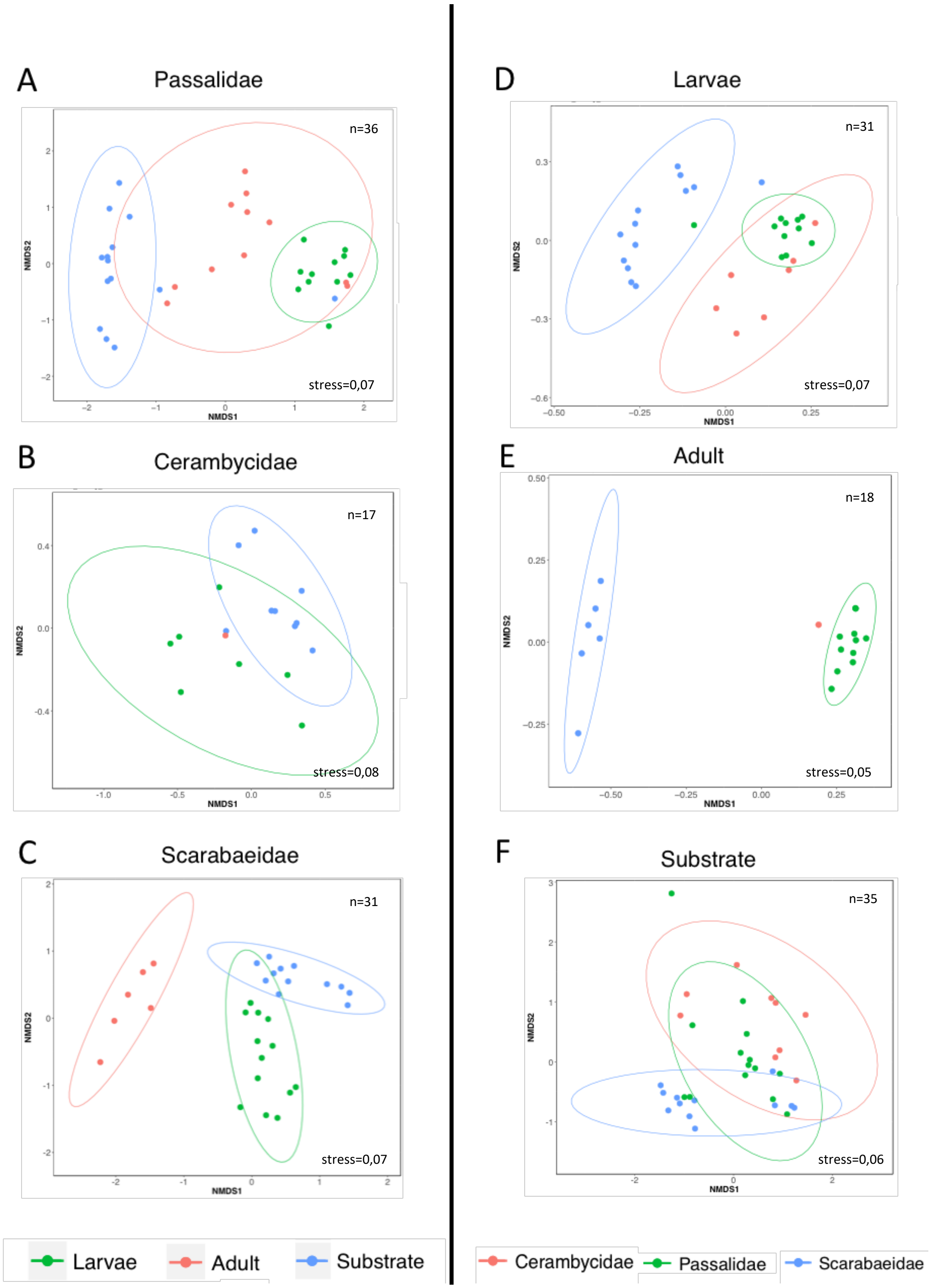

A
Passalidae
n=36
stress=0,07
NMDS1
NMDS2
B
Cerambycidae
n=17
stress=0,08
NMDS1
NMDS2
C
Scarabaeidae
n=31
stress=0,07
NMDS1
NMDS2
D
Larvae
n=31
stress=0,07
NMDS1
NMDS2
E
Adult
n=18
stress=0,05
NMDS1
NMDS2
F
Substrate
n=35
stress=0,06
NMDS1
NMDS2
Larvae
Adult
Substrate
Cerambycidae
Passalidae
Scarabaeidae

Figure 9. NMDS plot comparing the taxonomic composition of all beetle-associated samples colored by either beetle family (A,B and C) or sample type (D,E and F). Each dot on the plot corresponds to a different microbial community. Samples in A, B and C are colored according to sample type: larvae (green), adults (red) and substrate (blue) and samples D, E and F to beetle family: larvae (green), adults (red) and substrate (blue). A normal distribution for each group is shown by a continuous line. The NMDS stress value of each plot is shown in the bottom right corner along with the sample size (n) in the top right.

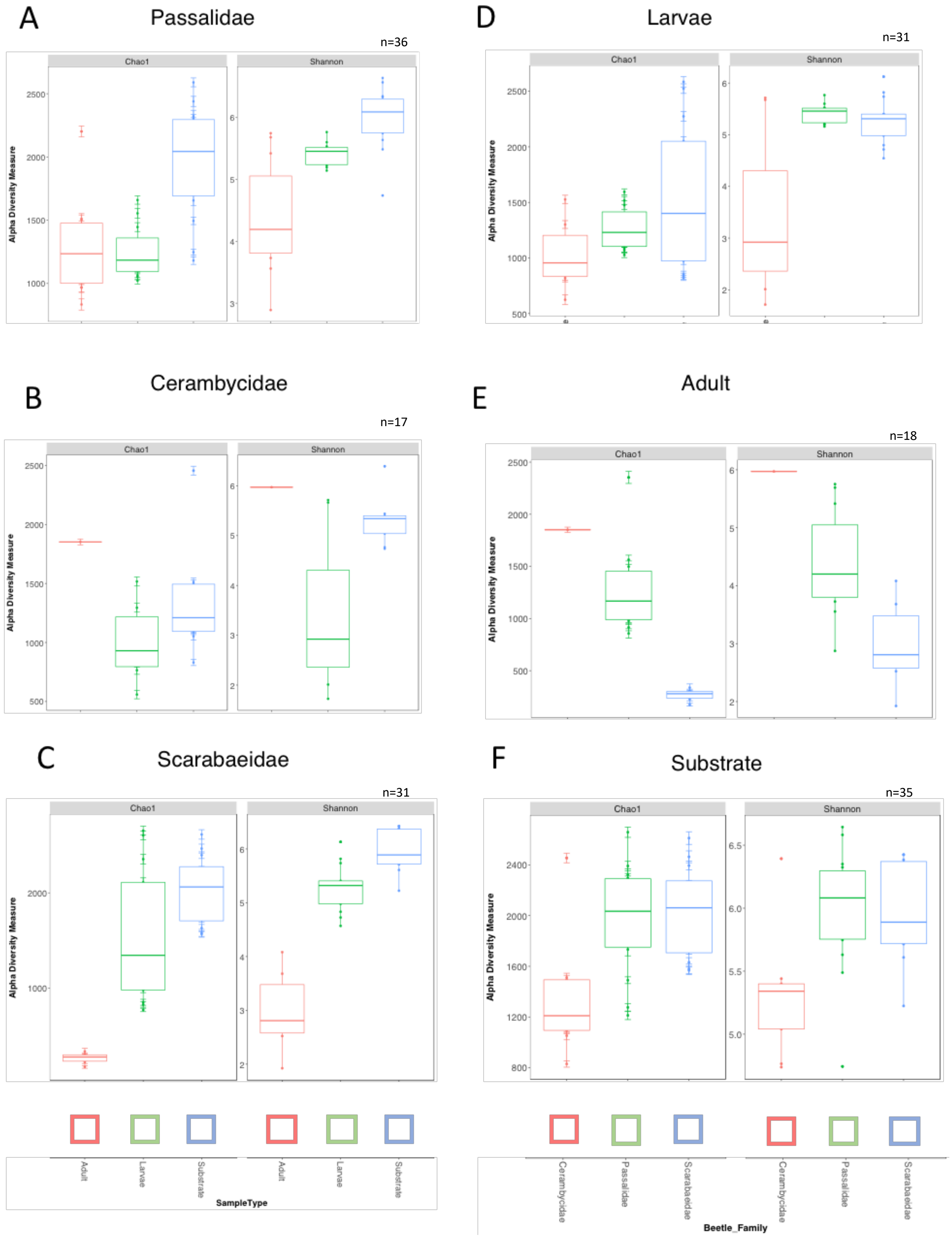
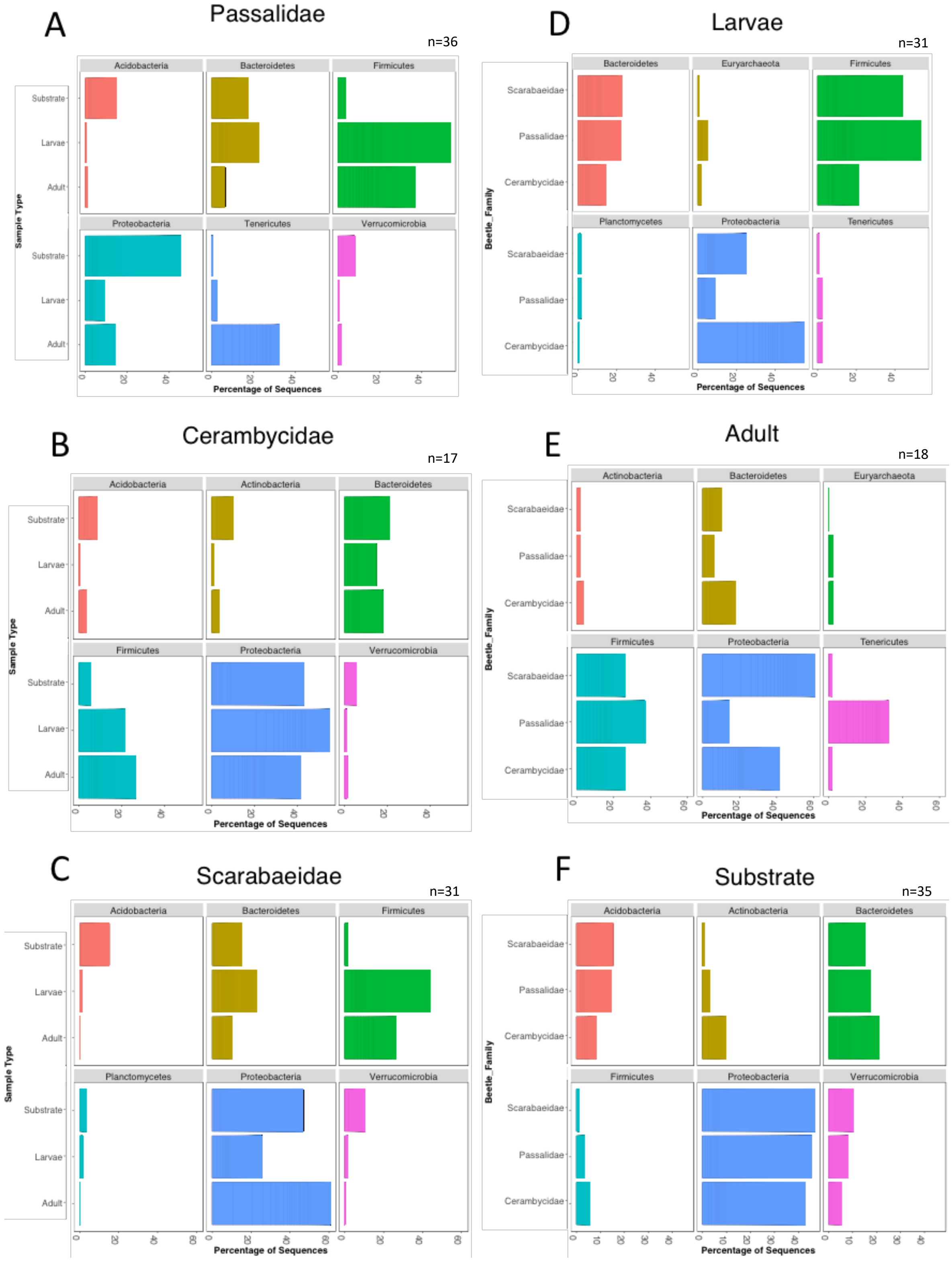

A
Passalidae
n=36
D
Larvae
n=31
B
Cerambycidae
n=17
E
Adult
n=18
C
Scarabaeidae
n=31
F
Substrate
n=35
Chao1
Shannon
Alpha Diversity Measure
Adult
Larvae
Substrate
SampleType
Cerambycidae
Passalidae
Scarabaeidae
Beetle_Family

Figure 10. Boxplots for Chao1 and Shannon alpha diversity index for all beetle associated samples by either beetle family (A,B and C) or sample type (D,E and F). Samples in A, B and C are colored according to sample type: larvae (green), adults (red) and substrate (blue) and samples D, E and F to beetle family: larvae (green), adults (red) and substrate (blue). The total sample size (n) for each plot is on the top right corner.



A
Passalidae
n=36
Acidobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Tenericutes
Verrucomicrobia
Sample Type
Substrate
Larvae
Adult
Percentage of Sequences
D
Larvae
n=31
Bacteroidetes
Euryarchaeota
Firmicutes
Planctomycetes
Proteobacteria
Tenericutes
Beetle_Family
Scarabaeidae
Passalidae
Cerambycidae
Percentage of Sequences
B
Cerambycidae
n=17
Acidobacteria
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
Sample Type
Substrate
Larvae
Adult
Percentage of Sequences
E
Adult
n=18
Actinobacteria
Bacteroidetes
Euryarchaeota
Firmicutes
Proteobacteria
Tenericutes
Beetle_Family
Scarabaeidae
Passalidae
Cerambycidae
Percentage of Sequences
C
Scarabaeidae
n=31
Acidobacteria
Bacteroidetes
Firmicutes
Planctomycetes
Proteobacteria
Verrucomicrobia
Sample Type
Substrate
Larvae
Adult
Percentage of Sequences
F
Substrate
n=35
Acidobacteria
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
Beetle_Family
Scarabaeidae
Passalidae
Cerambycidae
Percentage of Sequences

Figure 11. Relative abundance of the six most represented phyla for all beetle associated samples by either beetle family (A,B and C) or sample type (D,E and F). Samples in A, B and C are colored according to sample type: larvae (green), adults (red) and substrate (blue) and samples D, E and F to beetle family: larvae (green), adults (red) and substrate (blue). The total sample size (n) for each plot is on the top right corner.

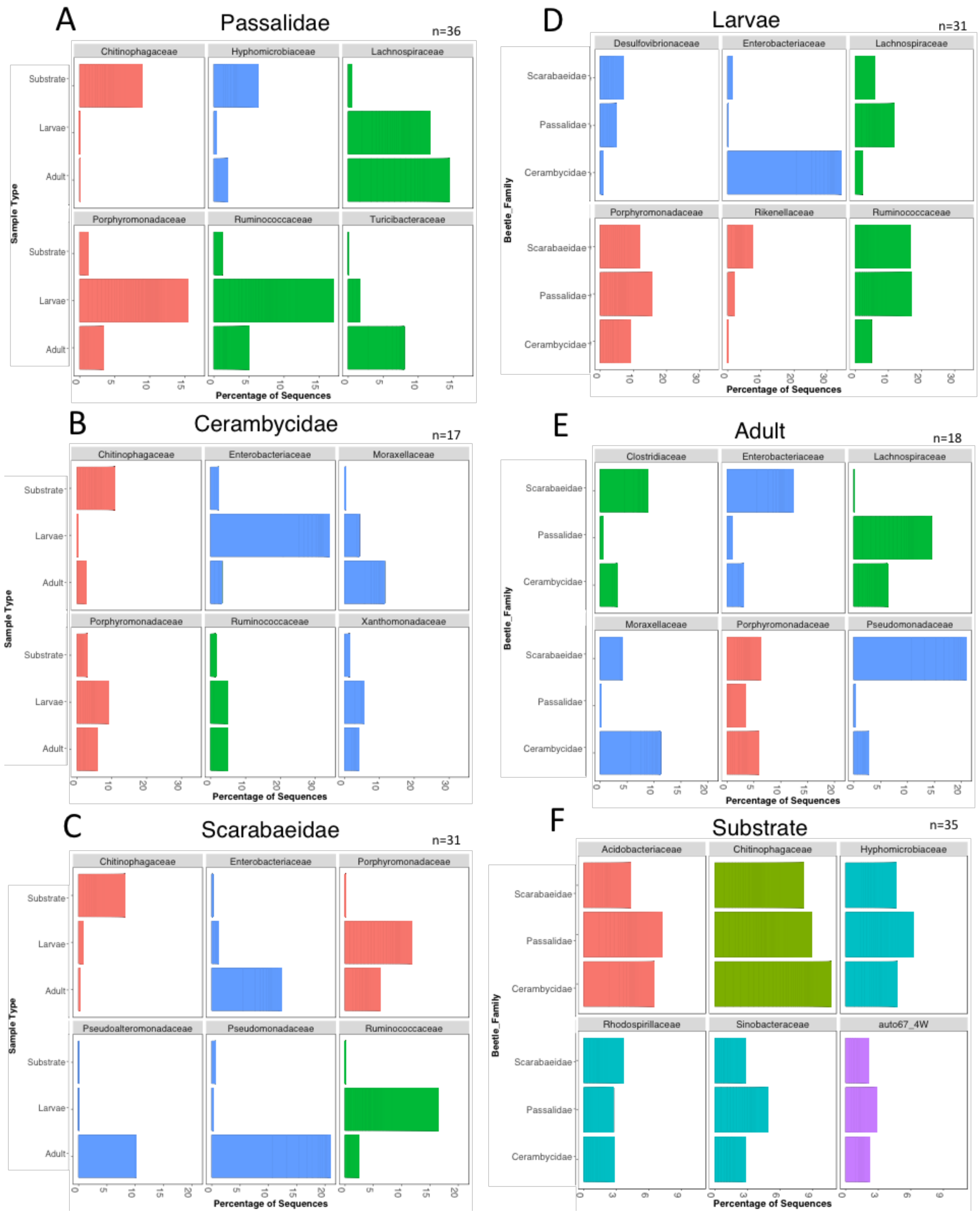
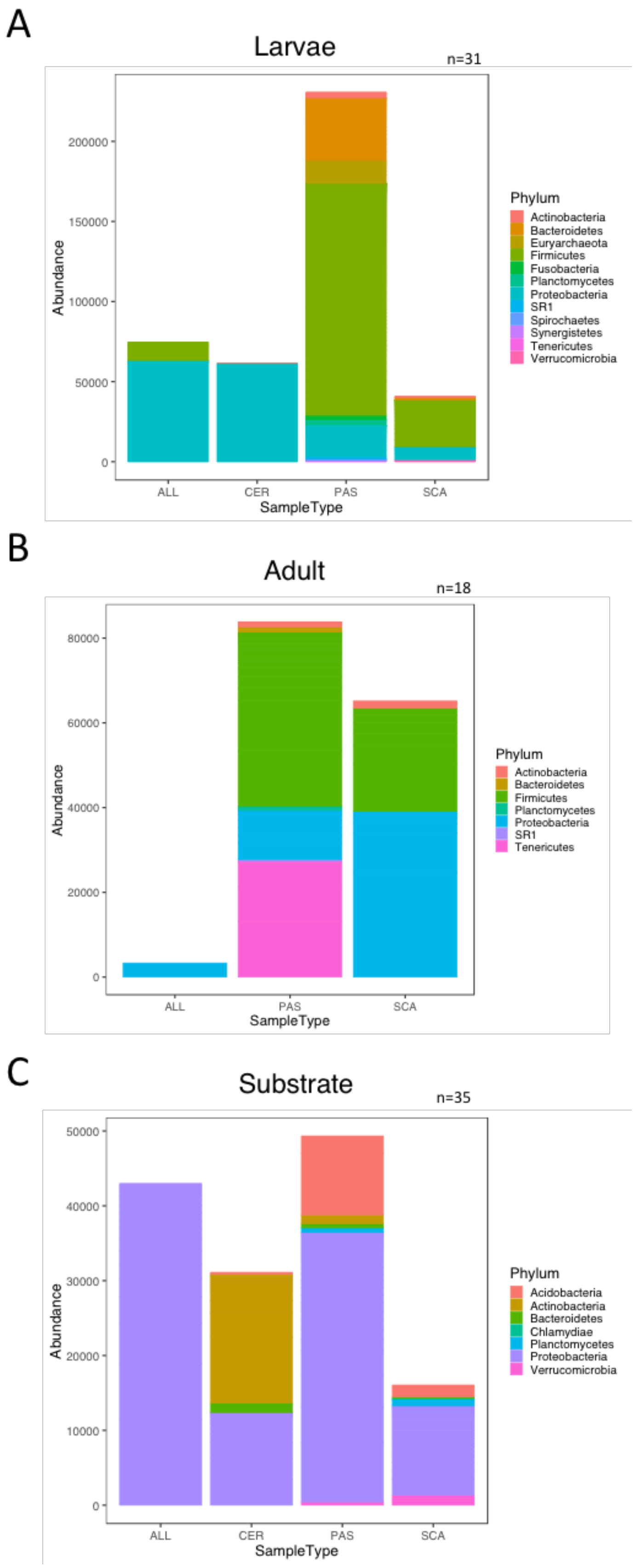

A
Passalidae
n=36
Chitinophagaceae
Hyphomicrobiaceae
Lachnospiraceae
Porphyromonadaceae
Ruminococcaceae
Turicibacteraceae
Sample Type
Substrate
Larvae
Adult
Percentage of Sequences
B
Cerambycidae
n=17
Chitinophagaceae
Enterobacteriaceae
Moraxellaceae
Porphyromonadaceae
Ruminococcaceae
Xanthomonadaceae
C
Scarabaeidae
n=31
Chitinophagaceae
Enterobacteriaceae
Porphyromonadaceae
Pseudoalteromonadaceae
Pseudomonadaceae
Ruminococcaceae
D
Larvae
n=31
Desulfovibrionaceae
Enterobacteriaceae
Lachnospiraceae
Porphyromonadaceae
Rikenellaceae
Ruminococcaceae
Beetle_Family
Scarabaeidae
Passalidae
Cerambycidae
E
Adult
n=18
Clostridiaceae
Enterobacteriaceae
Lachnospiraceae
Moraxellaceae
Porphyromonadaceae
Pseudomonadaceae
F
Substrate
n=35
Acidobacteriaceae
Chitinophagaceae
Hyphomicrobiaceae
Rhodospirillaceae
Sinobacteraceae
auto67_4W

Figure 12. Relative abundance of the six most represented families for all beetle associated samples by either beetle family (A,B and C) or sample type (D,E and F). Samples in A, B and C are colored according to sample type: larvae (green), adults (red) and substrate (blue) and samples D, E and F to beetle family: larvae (green), adults (red) and substrate (blue). The total sample size (n) for each plot is on the top right corner.



A
Larvae
n=31
200000
150000
100000
50000
0
Abundance
ALL
CER
PAS
SCA
SampleType
Phylum
Actinobacteria
Bacteroidetes
Euryarchaeota
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
SR1
Spirochaetes
Synergistetes
Tenericutes
Verrucomicrobia
B
Adult
n=18
80000
60000
40000
20000
0
Abundance
ALL
PAS
SCA
SampleType
Phylum
Actinobacteria
Bacteroidetes
Firmicutes
Planctomycetes
Proteobacteria
SR1
Tenericutes
C
Substrate
n=35
50000
40000
30000
20000
10000
0
Abundance
ALL
CER
PAS
SCA
SampleType
Phylum
Acidobacteria
Actinobacteria
Bacteroidetes
Chlamydiae
Planctomycetes
Proteobacteria
Verrucomicrobia

Figure 13. Abundance of core microbiome groups in at least 90% of larval samples by beetle family. A) Larvae B) Adult and C) Substrate. The total sample size (n) for each plot is on the top right corner. ALL= Every sample, CER= Cerambycidae, PAS= Passalidae, SCA= Scarabaeidae.

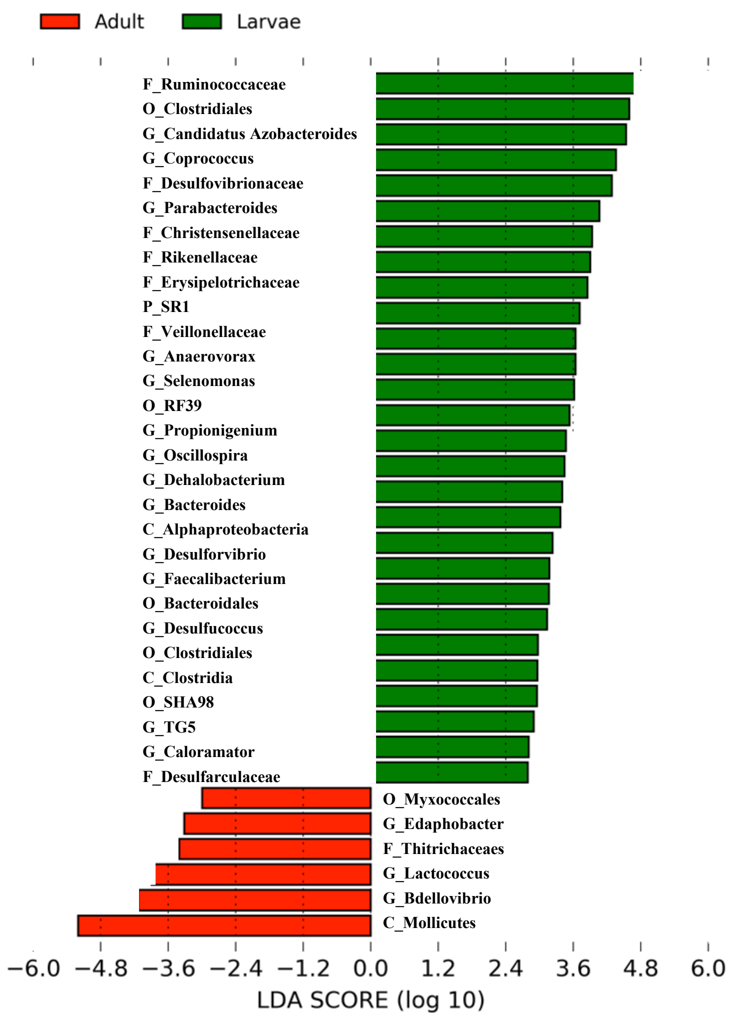


Figure 14. OTUs that best characterize differences between Pasalid adults and larvae LDA Effect Size (LEfSe) algorithm on genus level OTU tables to determine significant taxa. Labels for each OTU correspond to the lower taxonomy classification identified. P=phylum, C=Class, O=Order, F=Family, G=Genus.

6.2 Abundance of carbohydrate active enzymes from the Ruminococcaceae family.

Statistics for all four metagenomes downloaded from IMG and included in this analysis are displayed in Table 21. Amongst them, a total of 2 391 673 contigs were assembled using the IMG platform, resulting in 1519 Megabases (Mb) of sequence data. The annotated data resulted in a total of 3 462 736 protein coding genes, out of which 1,4 million had a known function assigned to them.

Table 21. Sample statistics for whole metagenome shotgun sequences downloaded from IMG.

	Adult Pool (Samples 3 & 4)	Substrate Pool (Samples 3 & 4)	Larvae 3	Larvae 4
	Sequences			
Percentage Assembled	0,64	100	0,81	100
Number of sequences	305261	380485	1112103	593824
Number of bases	142922252	170323120	804003313	402413667
GC count	46653124	94365726	344063477	175694446
	Genes			
RNA genes	774	4605	15934	2187
Protein coding genes	424257	428463	1786624	823392
With Product Name	24077	188691	709600	443694

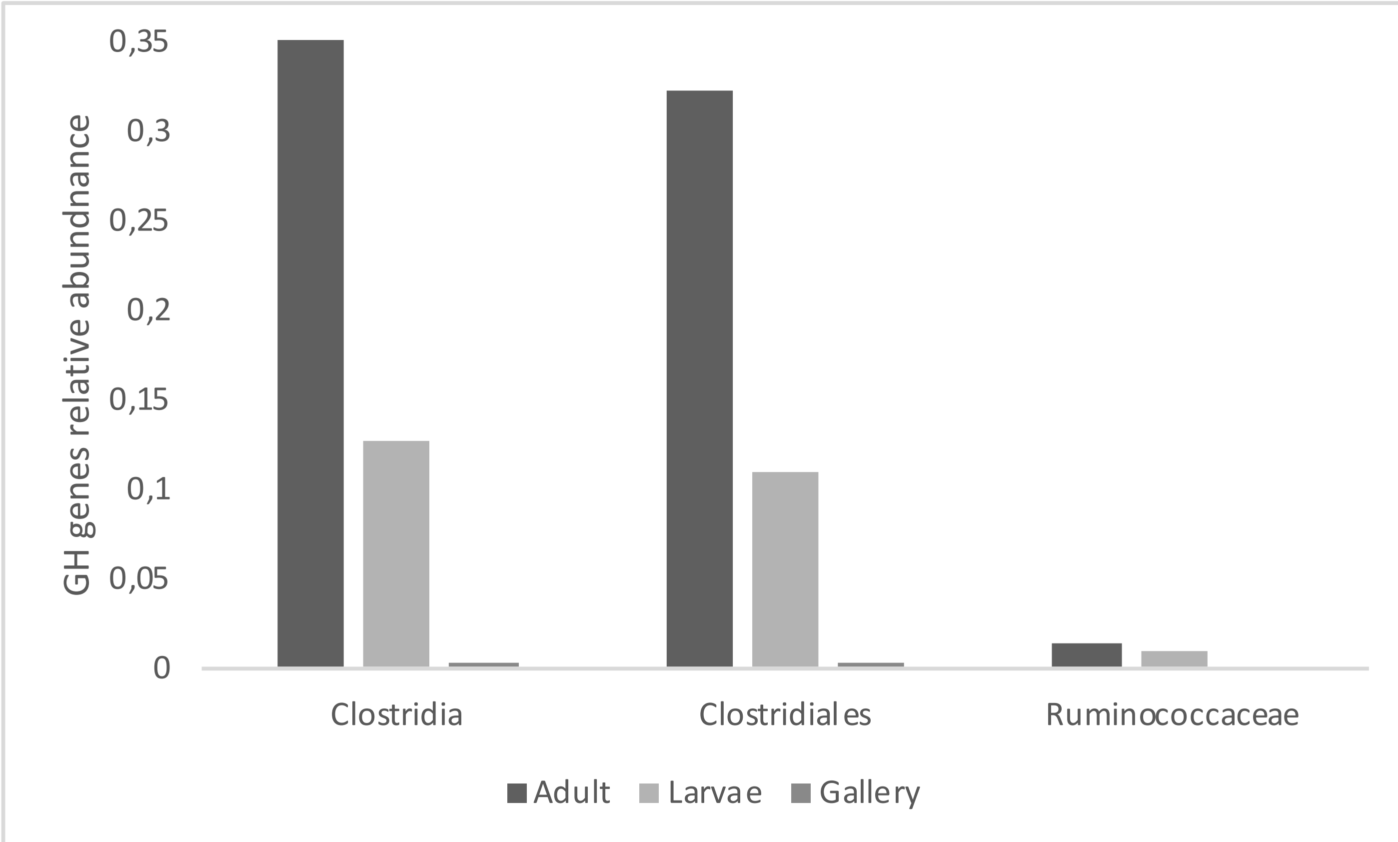


Figure 15. Phylogenetic distribution of Glycosil Hydrolases (GH) in genes from A) adult gut (n=98), B) substrate (n=942) and C) larval gut (n=5468) metagenomes by taxonomic rank (From class to family). Numbers indicate the relative abundance against the total number of pfam hits of each metagenome.

.

First, the carbohydrate metabolism capacity of the Ruminococcaceae sequences in the Passalid metagenomes were analyzed. The total amount of glycosyl hydrolase (GH) genes in the metagenomes was extracted to assign a taxonomic rank, afterwards only genes from the Ruminococcaceae family or related taxons such as order Clostridiales or class Clostridia were selected for downstream analysis. The number of GH genes that belong to any of these taxonomic categories should be directly proportional to the capacity of that group to metabolize carbohydrates. In the adult metagenome (Figure 15), most of the genes were assigned to the Clostridia and Clostridiales, while in the substrate GH genes were mostly assigned to the Acidobacteria and Proteobacteria and, in a lesser extent, to Ruminococcaceae related taxons. The reduced number of Clostridia, Clostridiales and Ruminococcaceae genes suggest a lesser involvement in carbohydrate breakdown within the substrate. Larvae samples had a higher abundance of GH genes. Contrary to the 16S rRNA gene results where Firmicutes was the most abundant phylum, most of the GH genes were assigned to the Bacteroidetes phyla. Around 32% of all sequences, still an important proportion of the genes, were assigned to Firmicutes (861). Within them, more than half of the genes are assigned to Clostridia. Most of the genes assigned to Clostridia were also assigned to Clostridiales. Finally, within the Clostridiales, the proportion of genes assigned to the Ruminococcaceae family was similar between larvae and adults.

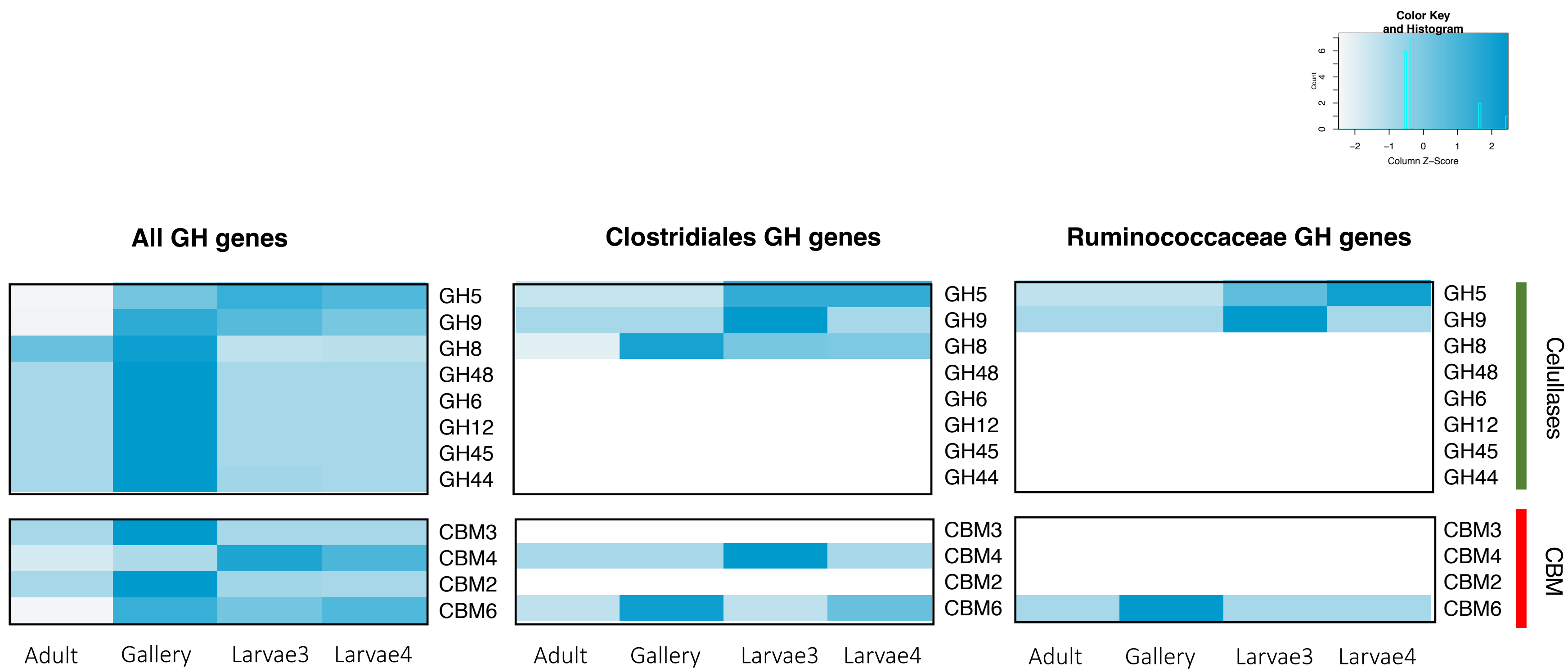


Figure 16. Abundance of GH and CBM putative genes in Passalid beetle metagenomes. Heatmaps are organized from the left (All GH genes in the metagenomes) to the right (only GH genes that were assigned to the Ruminococcaceae family). Data was normalized to the total pfam count of each metagenome.

After evaluating their overall carbohydrate degrading capacity, the next step was to look specifically for cellulase-coding genes. Therefore, a list of cellulase and CBM gene hits present in each metagenome is shown in Figure 16. Sequences assigned to order Clostridiales and family Ruminococcaceae were selected and plotted as heatmaps to indicate which cellulases were most prevalent in these taxonomic groups. With the exception of CBM4, every cellulase and CBM gene were present in the substrate compared to the other samples. The adult metagenome, containing fewer sequences, had a high proportion of GH8, CBM3 and CBM2 genes. Both larval metagenomes showed similar traits with an elevated number of hits belonging to GH5 (one of the most studied group of cellulases), and GH9 genes. Moreover, CBM4, underrepresented in substrate and adult metagenomes, is more commonly found in the larval metagenomes. CBM6 is found in both larvae, but also in the associated substrate. The gene fraction of GH assigned to Clostridiales had GH5, GH8 and GH9 genes, but only CBM4 and CBM6 were found. Also, most of these Clostridiales genes were present in the larval metagenomes. GH genes in the Ruminococcaceae family were exclusively from the GH5 and GH9 family, while CBM6 was the only CBM found associated with this group through the pfam search. Metagenomes of the Passalidae family are abundant in Ruminococcaceae sequences. Even more, Ruminococcaceae sequences present in those metagenomes can code for cellulases. These results, together with the MAG collection obtained in previous work [51], suggest that this family is as a strong candidate for further testing for cellulase activity, as follows next.

6.3 Phylogenomic and pangenomic analyses of Ruminococcaceae and Clostridiales related MAGs.

The MAGs collection was initially composed by a total of 768 Bins that were recovered from the metagenomic data. From those, 101 were selected for this study because they exhibited: 1) >70% completion 2) <10% redundancy 3) At least 1Mb of length. A preliminary taxonomic assignment was done by comparing the taxonomy assignment of the single copy core gene (sccg) collection in each bin with a known database (Table 22). From this strategy, a total of 24 MAGs were assigned as Ruminococcaceae, 4 as Lachnospiraceae and 18 could not be classified further than order Clostridiales. After confirming their taxonomy, MAGs classified as Clostridiales were selected for phylogenomic analysis. Table 23 shows the main statistics for each of the 46 MAGs that were preliminaryly identified as related to Clostridiales.

Table 22. Preliminary taxonomy of Clostridiales related Bins based on SCCG assignment.

MAG_code	Ruminococcaceae sccg count	Lachnospiraceae sccg count	Clostridiales sccg count	Total sccg count	Ratio R/Total	Ratio L/Total	Ratio C/Total	Preliminary taxonomy
bin.357	33	0	37	49	0,67	0,00	0,76	Ruminococcaceae
bin.86	33	2	46	61	0,54	0,03	0,75	Ruminococcaceae
bin.174	35	8	52	70	0,50	0,11	0,74	Ruminococcaceae
bin.133	25	3	29	53	0,47	0,06	0,55	Ruminococcaceae
bin.58	26	1	38	59	0,44	0,02	0,64	Ruminococcaceae
bin.519	17	1	27	40	0,43	0,03	0,68	Ruminococcaceae
bin.195	20	8	33	48	0,42	0,17	0,69	Ruminococcaceae
bin.206	17	2	24	42	0,40	0,05	0,57	Ruminococcaceae
bin.503	28	3	39	79	0,35	0,04	0,49	Ruminococcaceae
bin.317	17	4	26	48	0,35	0,08	0,54	Ruminococcaceae
bin.134	13	1	25	39	0,33	0,03	0,64	Ruminococcaceae
bin.340	22	4	40	72	0,31	0,06	0,56	Ruminococcaceae
bin.511	11	8	29	42	0,26	0,19	0,69	Ruminococcaceae
bin.505	13	6	29	51	0,25	0,12	0,57	Ruminococcaceae
bin.184	15	4	24	68	0,22	0,06	0,35	Ruminococcaceae
bin.440	11	4	31	52	0,21	0,08	0,60	Ruminococcaceae
bin.24	11	4	19	52	0,21	0,08	0,37	Ruminococcaceae
bin.123	11	0	31	53	0,21	0,00	0,58	Ruminococcaceae
bin.470	11	6	29	53	0,21	0,11	0,55	Ruminococcaceae
bin.481	12	0	30	58	0,21	0,00	0,52	Ruminococcaceae
bin.16	13	2	18	63	0,21	0,03	0,29	Ruminococcaceae

bin.89	12	1	26	59	0,20	0,02	0,44	Ruminococcaceae
bin.41	15	9	43	76	0,20	0,12	0,57	Ruminococcaceae
bin.14	12	1	23	61	0,20	0,02	0,38	Ruminococcaceae
bin.72	9	11	27	51	0,18	0,22	0,53	Lachnospiraceae
bin.301	5	35	52	55	0,09	0,64	0,95	Lachnospiraceae
bin.193	4	20	27	51	0,08	0,39	0,53	Lachnospiraceae
bin.272	1	21	37	54	0,02	0,39	0,69	Lachnospiraceae
bin.316	0	2	22	62	0,00	0,03	0,35	Unclassified Clostridiales
bin.127	7	1	23	48	0,15	0,02	0,48	Unclassified Clostridiales
bin.161	9	0	23	70	0,13	0,00	0,33	Unclassified Clostridiales
bin.361	8	2	26	63	0,13	0,03	0,41	Unclassified Clostridiales
bin.526	11	6	26	60	0,18	0,10	0,43	Unclassified Clostridiales
bin.4	5	3	21	61	0,08	0,05	0,34	Unclassified Clostridiales
bin.223	10	6	30	55	0,18	0,11	0,55	Unclassified Clostridiales
bin.96	10	3	28	55	0,18	0,05	0,51	Unclassified Clostridiales
bin.322	10	4	43	63	0,16	0,06	0,68	Unclassified Clostridiales
bin.485	6	4	40	79	0,08	0,05	0,51	Unclassified Clostridiales
bin.518	3	6	44	58	0,05	0,10	0,76	Unclassified Clostridiales
bin.405	3	1	61	88	0,03	0,01	0,69	Unclassified Clostridiales
bin.406	1	1	40	56	0,02	0,02	0,71	Unclassified Clostridiales
bin.64	1	1	68	82	0,01	0,01	0,83	Unclassified Clostridiales
bin.442	1	4	64	86	0,01	0,05	0,74	Unclassified Clostridiales
bin.330	1	3	69	95	0,01	0,03	0,73	Unclassified Clostridiales
bin.124	0	9	43	65	0,00	0,14	0,66	Unclassified Clostridiales
bin.552	0	6	53	88	0,00	0,07	0,60	Unclassified Clostridiales

Table 23. Sample statistics for best recovered Bins assigned to the Clostridiales from a list of 101 MAGs generated from the Passalid metagenome data.

Bin_code	Total length	Number of contigs	N50	GC_content	Percent completion	Percent redundancy
bin.552.fa	2254344	143	23041	33	99	4
bin.340.fa	1785043	152	17065	51	99	3
bin.127.fa	2762037	90	55910	51	97	4
bin.470.fa	2477736	159	20932	50	97	2
bin.505.fa	2247530	178	17326	42	97	4
bin.322.fa	2224147	222	13019	35	97	1
bin.330.fa_1	1501955	174	11303	26	96	6
bin.124.fa_1	2792969	247	17612	35	96	6
bin.485.fa	1824596	187	13368	32	96	4
bin.41.fa	2140966	150	21377	58	95	1
bin.405.fa	1727392	110	25836	27	95	6
bin.317.fa_1	1478132	129	15668	41	95	7
bin.184.fa	2122613	212	13166	53	94	1
bin.195.fa	2071625	131	25436	41	94	1
bin.526.fa	2034256	139	24169	47	93	6
bin.123.fa_1	1928882	160	18352	38	93	2
bin.316.fa	1264829	76	21964	29	93	3
bin.174.fa	3168773	265	15063	49	92	3
bin.503.fa	2145093	153	18626	54	92	1
bin.361.fa	1817726	157	14407	57	92	1
bin.4.fa	2107329	99	36416	53	91	2
bin.511.fa	2231954	178	15218	51	90	2
bin.64.fa_1	1442541	93	20121	28	88	4
bin.440.fa	1745623	175	13785	46	88	1

bin.518.fa_1	1669738	192	11340	37	87	4
bin.442.fa	1342311	61	35512	25	87	3
bin.72.fa	2026460	266	8858	48	86	4
bin.89.fa	3267142	65	90990	50	86	2
bin.58.fa	1353409	196	7932	49	85	1
bin.14.fa	1573518	158	12697	50	83	2
bin.161.fa	2601191	272	13020	60	82	2
bin.16.fa	2897163	381	9874	58	81	3
bin.481.fa	2634059	170	22640	52	81	1
bin.272.fa	2814037	148	24555	41	80	1
bin.86.fa	2251322	115	26149	55	78	2
bin.96.fa	2067989	67	54941	42	78	1
bin.223.fa	2836040	342	10232	54	76	5
bin.133.fa	1616156	103	20840	55	75	1
bin.206.fa	2648174	335	10027	39	73	2
bin.193.fa	2178890	175	16450	52	73	2
bin.519.fa	3346523	450	9284	43	73	1
bin.301.fa	1750208	136	25956	38	73	4
bin.406.fa_1	1281203	184	8045	28	73	3
bin.24.fa	1671390	148	15144	42	71	1
bin.134.fa	1218140	195	7104	50	71	1
bin.357.fa	1191913	150	9727	53	71	0

Phylogenomic analyses with the selected MAGs (Figure 17) showed that only 11 MAGs were identified as Ruminococcaceae because they cluster together in the branch with reference genomes. These 11 MAGs were preliminary identified as both Ruminococcaceae and Unclassified Clostridiales, showing discrepancies with the previous methodology. Only one MAG (Bin.301) clustered together with the reference genome *Ruminococcus gnavus* from the Lachnospiraceae family. The rest of the 32 MAGs did not group with Ruminococcaceae genomes, even more, Bin.316 diverged from all the other Clostridiales reference genomes and was placed in a separate branch from the overall phylogenetic tree. Every bin related to Clostridiales was found to be only in larvae as shown by the relative coverage column in Figure 17. Most MAGs were ensembled from sequences belonging to one of the two larvae metagenomes, but others such as Bin.357 or Bin.41 are evenly distributed in both larval samples.

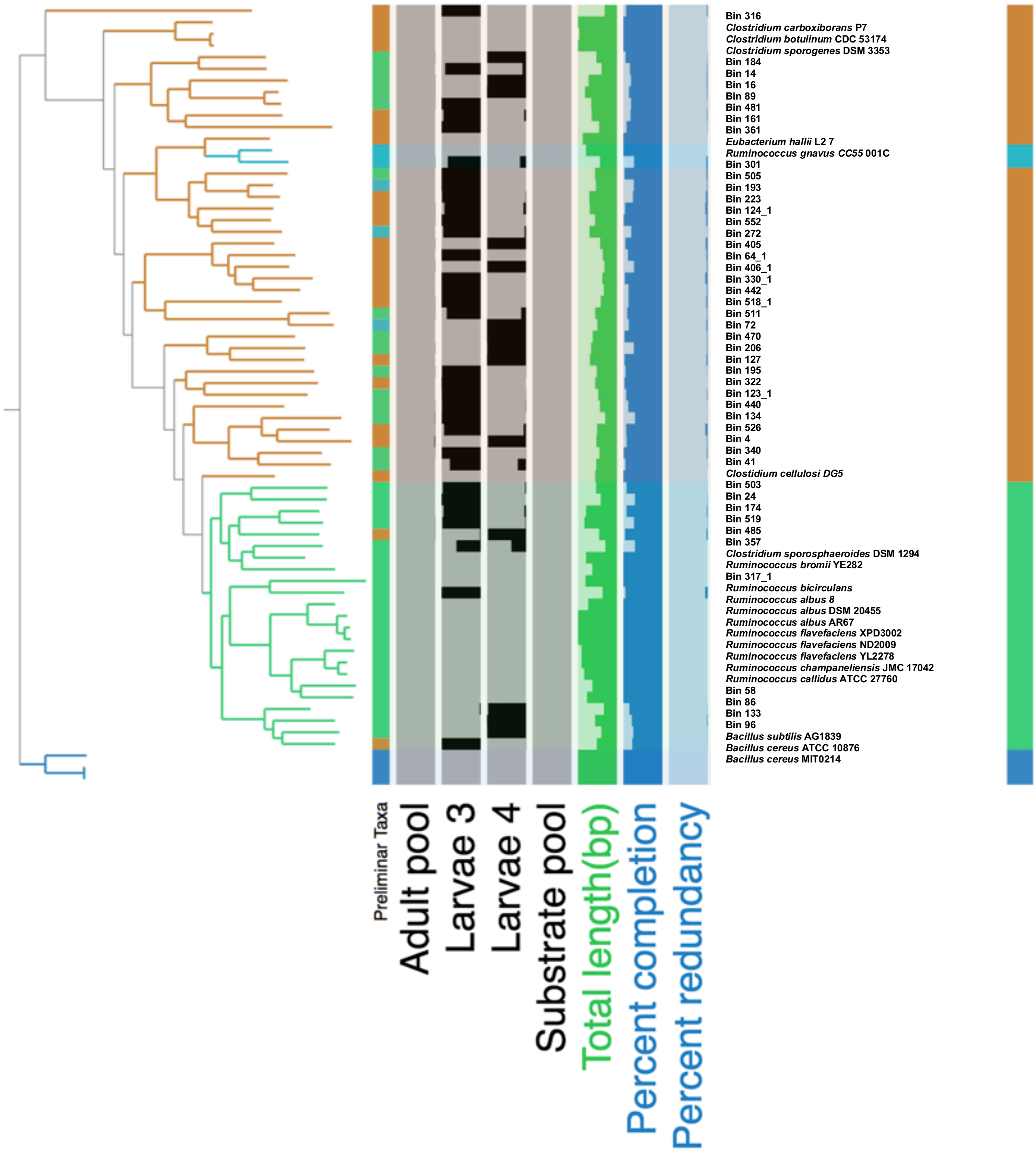

Bin 316
Clostridium carboxiborans P7
Clostridium botulinum CDC 53174
Clostridium sporogenes DSM 3353
Bin 184
Bin 14
Bin 16
Bin 89
Bin 481
Bin 161
Bin 361
Eubacterium hallii L2 7
Ruminococcus gnavus CC55 001C
Bin 301
Bin 505
Bin 193
Bin 223
Bin 124_1
Bin 552
Bin 272
Bin 405
Bin 64_1
Bin 406_1
Bin 330_1
Bin 442
Bin 518_1
Bin 511
Bin 72
Bin 470
Bin 206
Bin 127
Bin 195
Bin 322
Bin 123_1
Bin 440
Bin 134
Bin 526
Bin 4
Bin 340
Bin 41
Clostidium cellulosi DG5
Bin 503
Bin 24
Bin 174
Bin 519
Bin 485
Bin 357
Clostridium sporosphaeroides DSM 1294
Ruminococcus bromii YE282
Bin 317_1
Ruminococcus bicirculans
Ruminococcus albus 8
Ruminococcus albus DSM 20455
Ruminococcus albus AR67
Ruminococcus flavefaciens XPD3002
Ruminococcus flavefaciens ND2009
Ruminococcus flavefaciens YL2278
Ruminococcus champaneliensis JMC 17042
Ruminococcus callidus ATCC 27760
Bin 58
Bin 86
Bin 133
Bin 96
Bacillus subtilis AG1839
Bacillus cereus ATCC 10876
Bacillus cereus MIT0214
Preliminar Taxa
Adult pool
Larvae 3
Larvae 4
Substrate pool
Total length(bp)
Percent completion
Percent redundancy

Figure 17. Phylogenomic distribution of all good quality bins phylogenetically assigned to the Clostridiales class .The first bar represents the preliminary taxonomy of each MAG based on the Kbase database. The second to fifth bar shows the relative coverage of each bin in all four metagenomes. The sixth bar correspond to the total length of the MAGs. The seventh and eighth bars are the percent of completion and redundancy, respectively. And the ninth bar represents the different groups obtained by the phylogenomic analysis. The colors for the first and ninth bars are as follows: Green (Ruminococcaceae related bins), Orange (Other Clostridiales related bins), Light blue (Lachnospiraceae related bins), Blue (Root).

Every MAG assigned to the Ruminococcaceae family (green branches in the tree shown in Figure 17) can be separated into 6 clades (Figure 18) where the first group comprises only reference genomes. The second clade includes Bin.317 and one *Subduligranulum variable* genome. The third clade comprises four MAGs that cluster separately from any reference genome and would be candidates for novel species, similar to the fourth and fifth clades, which contain two and three MAGs respectively. The final sixth group contains one MAG, Bin.357 and two Ruminococcaceae genomes: *Ruminococcus bromii* and *Clostridium sporosphaeroides*. Although many MAGs are not clustering with reference genomes, they are still flanked by them, which means they share a common ancestor with the Ruminococcaceae taxa. These results provide evidence for genomes closely related to Ruminococacaceae organisms present in larvae, with the possibility of them being putative new species.

Table 24. Potential carbohydrate active enzymes related to cellulase activity present in MAGs ensembled from Passalid gut metagenomes. Each value in this table corresponds to the absolute number of hits in each genome. MAGs with cellulases are shaded in grey.

	Bin_24	Bin_133	Bin_485	Bin_357	Bin_58	Bin_317	Bin_519	Bin_86	Bin_503	Bin_96	Bin_174
Completeness	70,5	74,82	95,68	70,5	84,89	94,96	72,66	78,42	92,09	77,7	92,09
Length (Mb)	1,7	1,6	1,8	1,2	1,4	1,5	3,3	2,3	2,1	2,1	3,2
GH5	0	0	0	0	0	0	6	0	5	0	14
GH9	0	0	0	0	0	0	0	1	2	0	0
CBM4	0	0	0	0	0	0	3	0	0	0	0
CBM6	12	0	5	1	0	0	1	1	0	0	1
dockerin	6	0	0	0	0	0	17	0	0	0	20
cohesin	1	0	0	8	0	0	7	0	0	0	2

The pangenomic analysis performed with the Ruminococcaceae MAGs is showcased in Figure 18. The goal of this analysis was to identify protein clusters of similar functions common to each clade. Protein clusters were identified in the interactive interface of Anvi'o and marked (Figure 18). The functions associated to the core genome of every MAG and reference genome included a total of 19678 gene sequences distributed into 1289 protein clusters, from which 1186 were assigned to a specific COG function. Furthermore, genes identified as being shared only by MAGs or genomes from clades 2, 3 and 5 were selected and a table with the summary was generated. Genes from clade 4 were not shared among the two MAGs included in it. Appendix 3 contains every COG function identified from all protein clusters in clade 2, and which genomes (or MAGs) contained them. A total of 462 genes were detected in this group. Out of this total, 33 corresponded to category G or associated to "Carbohydrate metabolism and transport". The most remarkable sequence amongst them encodes for a Glycosil transferase 2 (GT2). These enzymes are involved in the biosynthesis of a variety of polysaccharides and could provide insights on how genomes of the clade 2 metabolize carbohydrates.

Appendix 4 includes the protein clusters shared among MAGs from clade 3. A total of 736 genes are included in this cluster and 54 of them belong to category G (Carbohydrate metabolism and transport) of the COG database. Genes such as 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, broad specificity phosphatase PhoE and 6-phosphofructokinase are present in at least 3 out of 4 of the MAGs included in the clade. All of them are centrally related to carbohydrate metabolism. On the other hand, appendix 5 contains shared genes from the three MAGs: bin.519, bin.174 and bin.485 in clade 5. Although none of the protein clusters for this clade are in bin.485, a total of 719 genes are found in both bin.519 and bin 174. From this 719 genes, 54 were assigned to category G in the COG database, including cellobiose phosphorylase involved in the metabolic reaction converting cellobiose into glucose and glucose-1-phosphate. Finally, the genes for the glycogen debranching enzyme (alpha-1,6-glucosidase) and glucoamylase (glucan-1,4-alpha-glucosidase) were found in the protein clusters in clade 5, both related to polysaccharide breakdown.

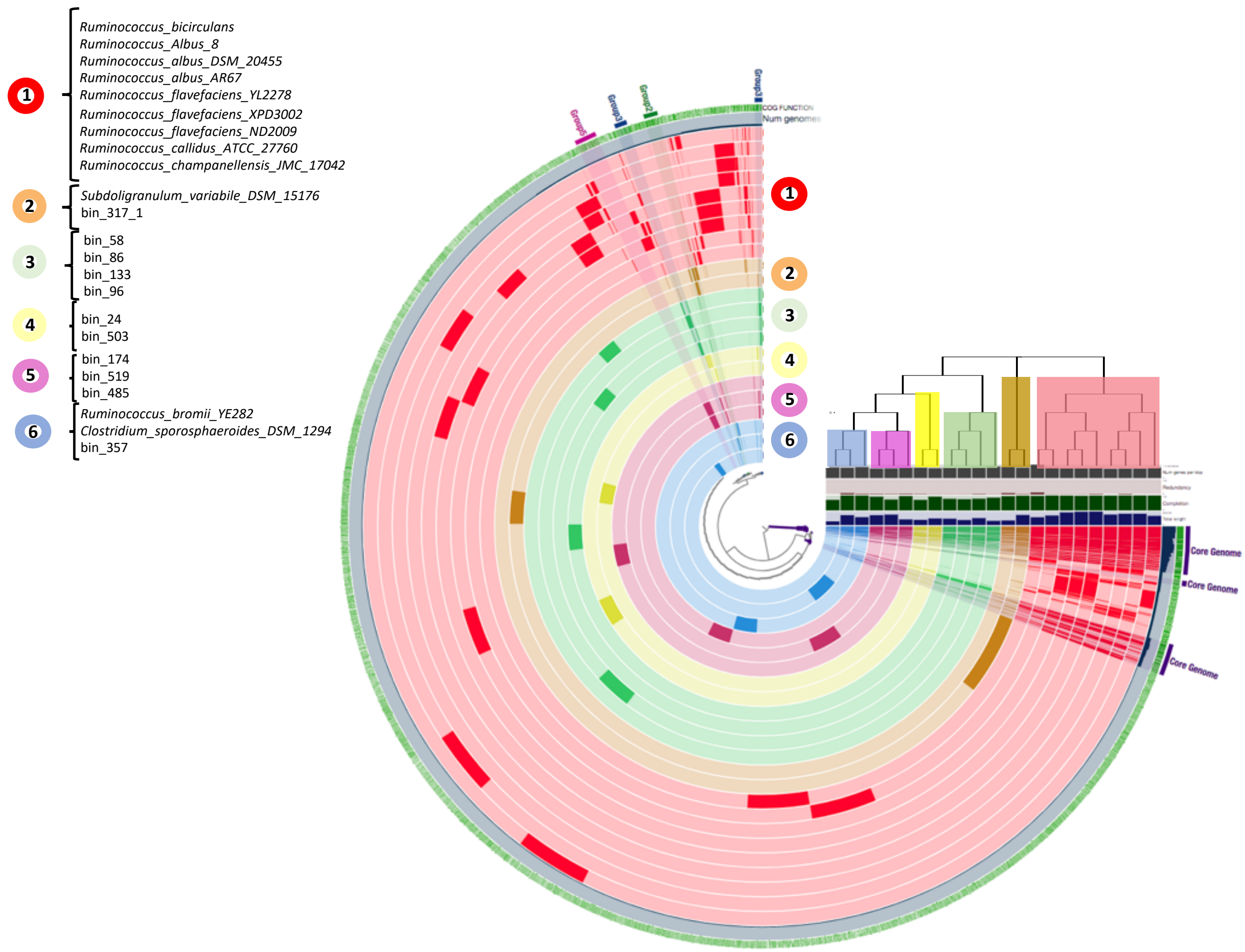


Figure 18. Pangenomic analysis of Ruminococcaceae MAGs recovered from larval metagenomes. Different colors indicate all different clades delimited by the maximum likelihood tree clustering.

Different collections of protein clusters associated to each clade are marked as shades on the side of the round bars. Collections of shared genes were generated for the core genome, and clades 2, 3 and 5.

Additionally, I searched for cellulose degrading enzymes specifically in every Ruminocococaceae MAG. The total amount of cellulase-related protein hits found in the Ruminococcaceae MAGs through HMM are displayed in Table 24. From all the GHs included in figure 16, only GH9 and GH5 were present in these MAGs; the latter being one of the most abundant and well described cellulase groups. GH5 hits were found in Bin 174, Bin 519 and Bin 503, while GH9 was found only in Bin 86 and Bin 503. Besides, the module CBM 6 was present in 6 out of 11 MAGs analyzed while CBM 4 was present only in Bin 519. On the other hand, CMB 2 and 3 were not identified in any of the Ruminococcaceae MAGs, opposed to the results in Figure 16. Other protein features found through the HMM search for cellulases were cohesin and dockering protein coding genes. Such protein domains, as discussed previously, are important components of cellulosomes. Bin 24, Bin 174 and Bin 519 had both dockering and cohesin domains while Bin 357 had only genes coding for scaffolding cohesion domains.

As stated here, Ruminococcaceae MAGs present in the Passalidae metagenomes encode for both carbohydrate metabolism proteins and specific cellulases. A complete profile of every carbohydrate active enzyme is provided in Appendix 6. Since these MAGs could be novel species with the potential to degrade cellulose, the final step of this work was to provide some insights in the possible ecological role of these organisms in the gut through the analysis of their genomes.

For this portion, the four MAGs with cellulase hits were selected: Bin 519, Bin 503, Bin 174 and Bin 86.

6.4 Genomic analysis of Ruminococcaceae MAGs related to cellulose degradation present in Passalidae metagenomes.

As discussed above, the genomes of the four Ruminococcaceae MAGs with cellulase hits were analyzed in more depth as shown in Figure 19. The coordinates of the four draft genomes indicate the position of the putative cellulases, dockering domains and CBMs (Table 24). Cellulases are distributed across the genome and they seem to be apart from each other with a few exceptions, particularly in Bin 174 and Bin 519. These last two draft genomes are the largest in terms of base pairs amongst all Ruminococcaceae MAGs identified in the analysis. Therefore, it is expected that they have the highest number of genes related to cellulose metabolism. As mentioned above, genomic regions where several cellulases and other cellulose metabolism genes were placed near each other were identified (Figure 20). Both Bin 519 and Bin 174 gene clusters included at least one dockering domain and several GH hydrolases. Also, in both cases, genes coding for permeases were found in the vicinity of these genes. In Bin 174 the gene cluster included a CBM domain plus the rest of the genes already mentioned above.

Table 25 shows the closest taxonomic assignments given by the RAST annotation server. For Bin 174 and Bin 519 the higher scores for both MAGs are the same: *Clostridium methylpentosum, C. leptum, Ethanoligenens harbinense and Anaerotruncus colihominis*. For Bin 86, the highest score was for *A. colihominis* and in second place *Clostridium thermocellum.* Bin 503 highest ID score was *E. harbinense* and *A. colihominis* was second. An overview of the metabolic potential of all four genomes is displayed in Figure 21. Genes are grouped according to different broad categories defined by the RAST sever. For every MAG, most of the genes assigned to a specific pathway were associated to carbohydrate and amino acid metabolism. Two of them had the highest number of genes associated to carbohydrates (Bin 519 and Bin 86) and the other two (Bin 503 and Bin 174) have a higher abundance of amino acid metabolism related genes. The third most abundant category is protein metabolism. Afterwards, several categories such as DNA, nucleotide, cofactors, cell wall and phosphorus metabolism were distributed at distinct proportions amongst the four MAGs. Secondary metabolism, chemotaxis and iron acquisition metabolism genes are absent from all four genomes suggesting that they are probably not present in the Ruminococcaceae group.

Table 25. Closest phylogenetic neighbors of bins 174, 503, 86 and 519 according to RAST-annotation server.

Genome ID	Score	Genome Name	Genome ID	Score	Genome Name
		Bin 174			**Bin 519**
537013.3	**500**	**Clostridium methylpentosum DSM 5476**	**537013.3**	**518**	**Clostridium methylpentosum DSM 5476**
428125.8	**480**	**Clostridium leptum DSM 753**	**428125.8**	**494**	**Clostridium leptum DSM 753**
663278.3	394	Ethanoligenens harbinense YUAN-3	**663278.3**	473	Ethanoligenens harbinense YUAN-3
663278.4	384	Ethanoligenens harbinense YUAN-3	**663278.4**	462	Ethanoligenens harbinense YUAN-3
445972.6	376	Anaerotruncus colihominis DSM 17241	**445972.6**	340	Anaerotruncus colihominis DSM 17241
		Bin 86			**Bin 503**
445972.6	**530**	**Anaerotruncus colihominis DSM 17241**	**663278.4**	**537**	**Ethanoligenens harbinense YUAN-3**
203119.11	**462**	**Clostridium thermocellum ATCC 27405**	**445972.6**	**372**	**Anaerotruncus colihominis DSM 17241**
350688.3	379	Alkaliphilus oremlandi oremlandii OhILAs	**203119.11**	313	Clostridium thermocellum ATCC 27405
663278.4	358	Ethanoligenens harbinense YUAN-3	**273068.3**	242	Thermoanaerobacter tengcongensis MB4
718255.3	324	Roseburia intestinalis XB6B4	**471871.7**	217	Clostridium sporogenes ATCC 15579

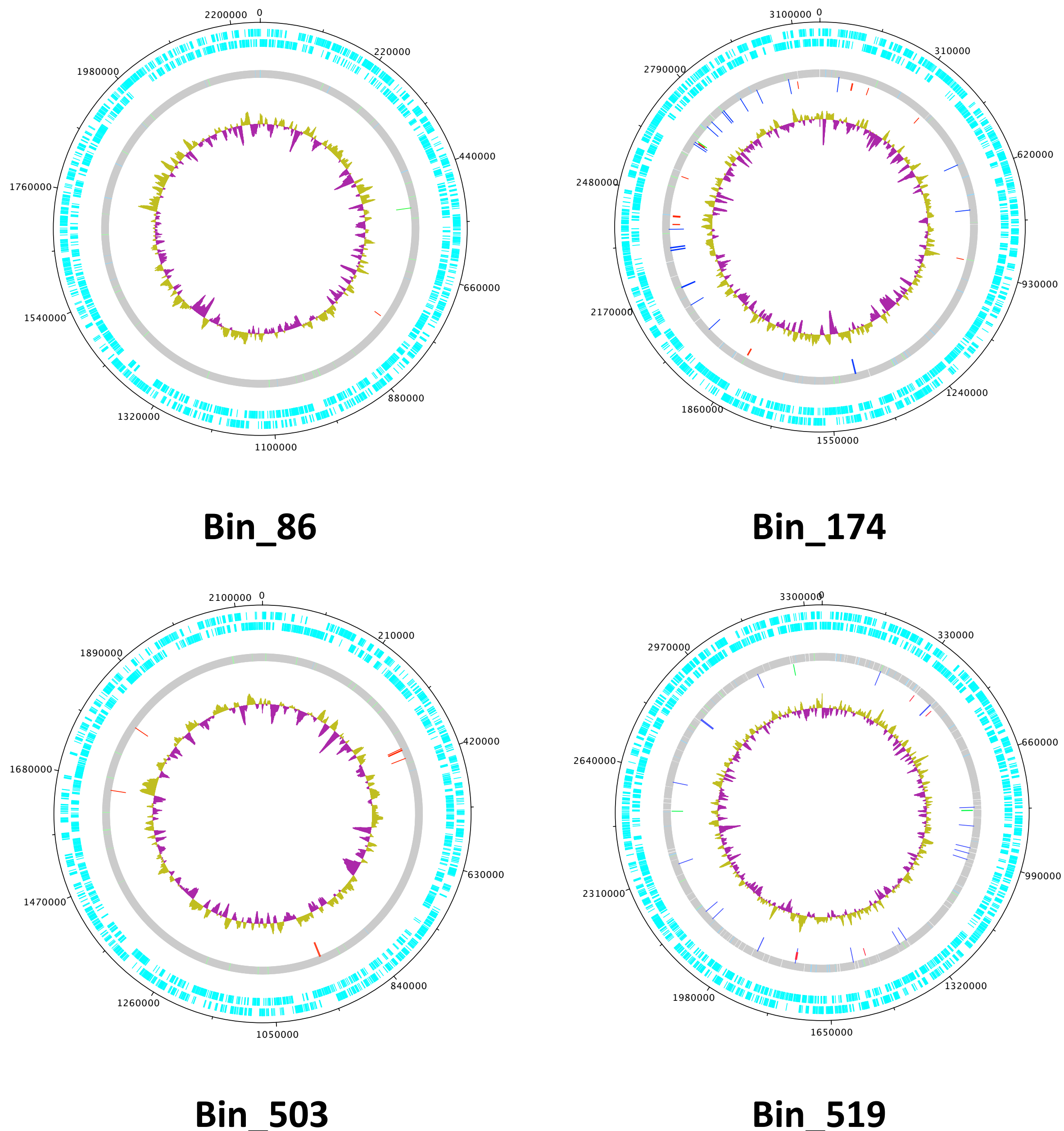


Figura 19. Ruminococcaceae MAGs with cellulase hits: The rings from the outside represent: 1. Coordinates in bp 2. Coding regions at the leading strand 3. Coding regions at the lagging strand 4 4. RNA sequences (Green) 5. Genes with hits for cellulases (Red), CBM (Green), and dockering domains (Blue) are represented with bars 6. GC plot

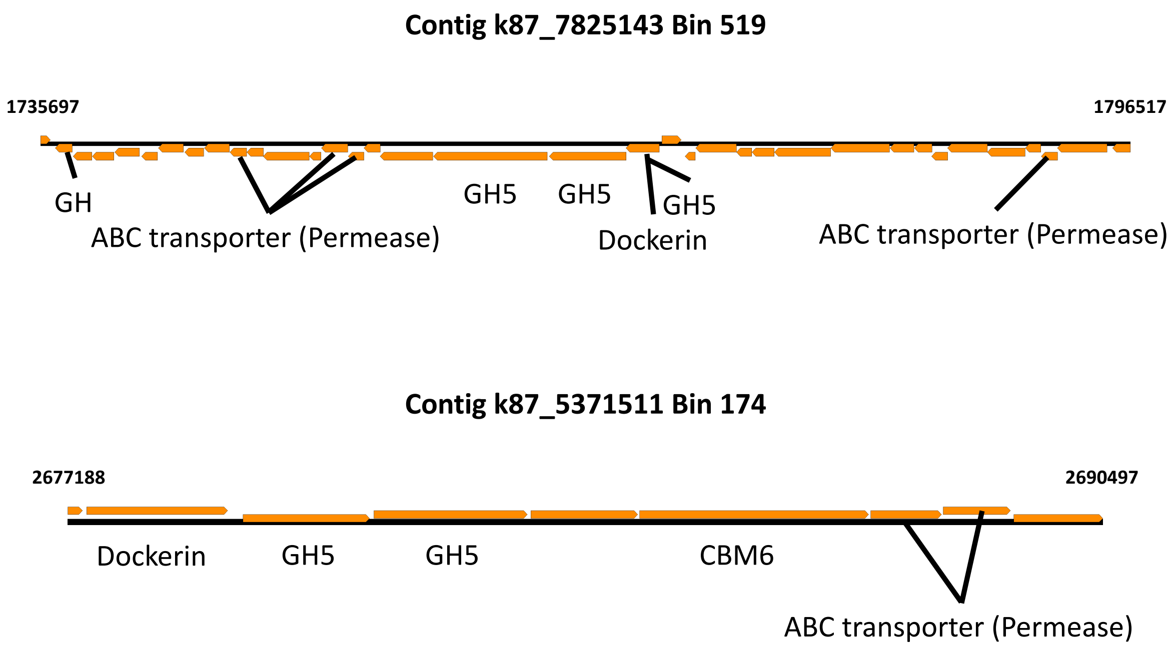


Figure 20. Contigs with gene clusters encoding cellulose metabolism related genes present in the selected MAGs. Orange frames represent each putative gene inside the contig. Orange frames above the black central line represent genes coded in the leading strand and below it in the lagging strand. The genes label GH5, CBM and dockering were discover through HMM analysis with dbCan. The rest of putative genes were assigned through the RAST annotation server.

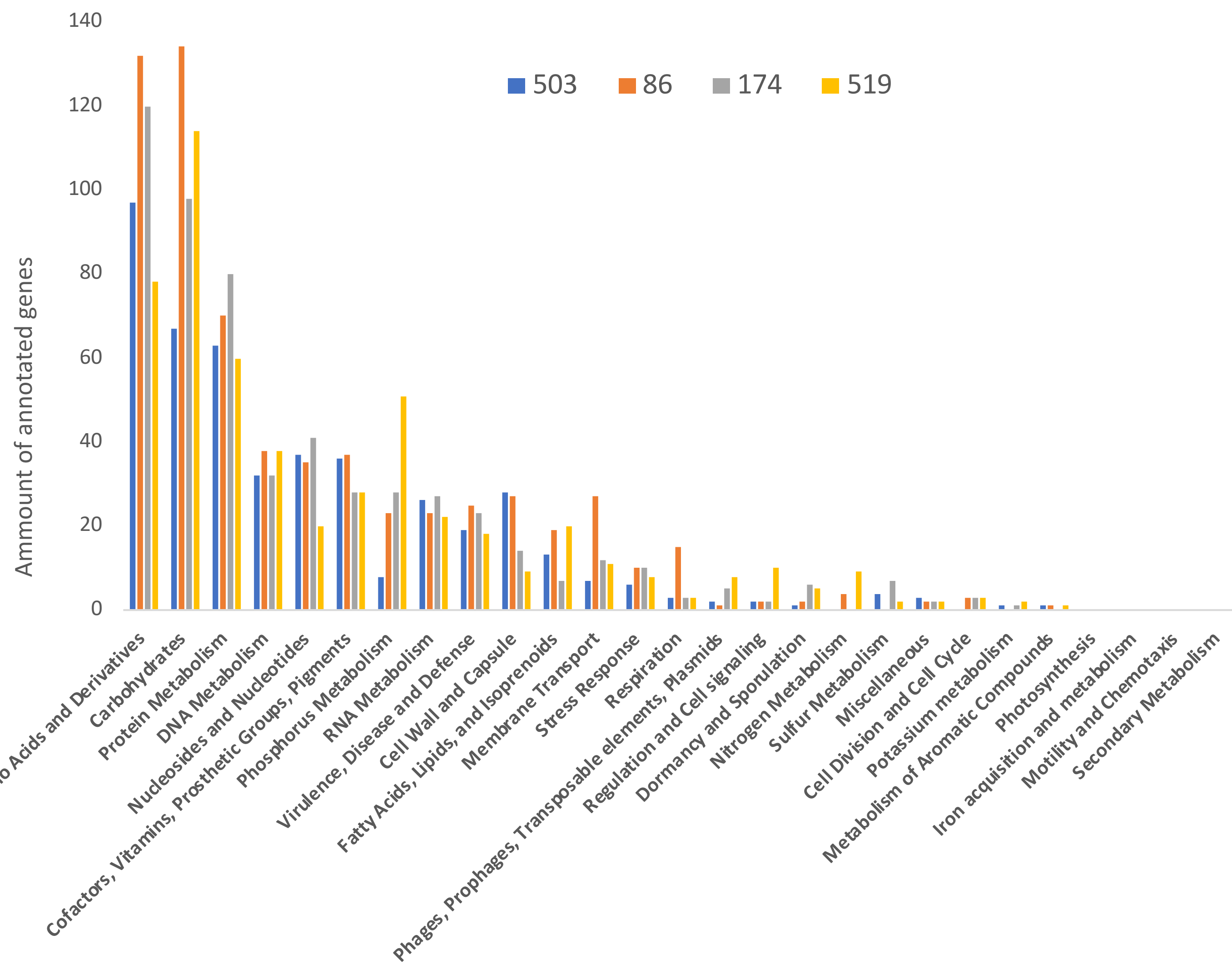


Figure 21. Abundance of genes associated to central metabolic pathways from MAGs with related to cellulose degrading potential. Function categories are sorted according to overall abundance.

An extensive search for glycolysis related genes in the four MAGs is shown in Figure 22. All enzymes required for the oxidation of glucose in order to obtain two molecules of pyruvate are present in every MAG. Protein encoding genes for the conversion of Glucose-6P to Fructose-1,6P and segmentation to molecules of Glyceraldehyde-3P and further conversion to pyruvate are all present in the contigs of the selected MAGs. Fermentation pathways for lactate, acetate and ethanol were also analyzed. The enzyme lactate dehydrogenase is present only in Bin.174. The enzyme aldehyde dehydrogenase is needed for the conversion of Acetyl-CoA to acetaldehyde, so that alcohol dehydrogenase can catalyze the conversion into ethanol. The first is absent from the MAGs, but alcohol dehydrogenase is present in all four of them. The pathway of acetate biosynthesis first needs to transform the acetyl-coA molecule into acetyl-P, catalyzed by the phosphate acetyltransferase; next acetyl-P is turn into acetate by the acetate kinase enzyme. Neither of the genomes had the phosphate acetyltransferase coding gene, however, the second enzyme needed to complete this process was present in all MAGs. If this pathway was complete in any of the four MAGs, the putative production of acetate could lead to the accumulation of this SCFA in the Passalid Gut. Finally, if other energy sources were scarce in the ecosystem, the four genomes are in full capacity to use the same acetate as an energy source. This way it could enter the central metabolism pathways through its conversion into acetyl-CoA by the acetyl-CoA synthase.

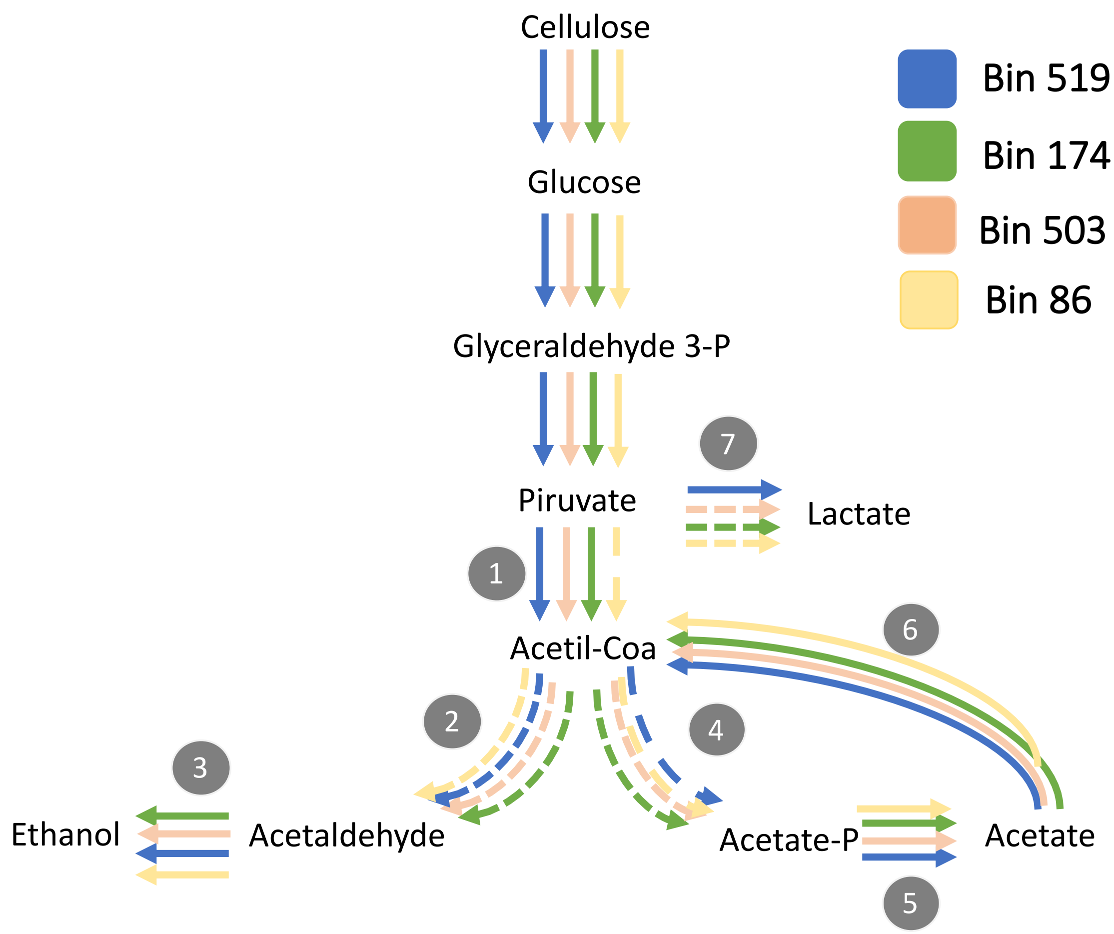


Figure 22. Presence of genes related with glycolysis and linked pathways from selected MAGs related to cellulose degrading potential. Continuous lines indicate that the enzyme is coded in the specific MAG, dashed lines indicate that the enzyme is absent. Enzymes are represented with a number as follows: 1. Pyruvate ferredoxin oxidoreductase (K00169) 2. Aldehyde dehydrogenase (K00128, K00129) 3. Alcohol dehydrogenase (K00001, K00002) 4. Phosphate acetyltransferase (K00625, K13788) 5. Acetate kinase (K00925) 6. Acetyl-CoA synthase (K01895) 7. Lactate dehydrogenase (K00016).

7. Discussion

Herbivores and xylophagous organisms are essential to maintain the balance of the carbon cycle in every ecosystem[94]. The degradation of cellulose and other recalcitrant carbon sources varies depending on the organism. Nonetheless, organisms that perform this process in nature can be divided in two groups, based on whether they produce their own cellulases or rely on symbiotic relationships with microorganisms that can express these enzymes [34]. Evidence from previous research suggested that the Passalidae beetle family is enriched in microorganisms capable of degrading cellulose. The work done by Vargas analyzed the uncultured microbiota of these beetles, but was limited to one location and samples from the Passalid family only[51]. Here, I present evidence on how sample type drives most of the differences in Passalid beetles microbiome, regardless of sample location and beetle family.

In order to accomplish this, first the microbial communities of different beetle samples were analyzed through 16S rRNA gene amplicon libraries. Besides sample type, both host family and geographic location seem to influence the community structure in these insects. Beetles were collected in three different ecosystems to evaluate a possible effect of each site (climate, plant availability, altitude, among others) over the composition of microbial communities. The Coco Island National park has been apart from the continent since its formation, about 2 million years ago[95]. Hence, beetles that inhabit the island have been separated from the mainland for a significant amount of time. The influence of taxonomic relatedness, location and sample type as factors that affect bacterial community composition is not unexpected as it has been described previously. Yun and colleagues found that the host environmental habitat, diet, phylogeny and the developmental stage are the main drivers for the differences in gut bacterial communities amongst insects [96,97].

Yun also discusses that one of the main differences amongst insect gut communities is the presence of anaerobic bacteria[96]. In this study, Firmicutes is the more abundant phylum detected in larval and adult guts, while Proteobacteria is the more abundant in the substrate. These groups seem to be dominant phyla and drive most of the differences within sample types. Other groups such as Bacteroidetes have fewer variations in abundance for every sample type. One discrepancy over Yun's findings was the notion that the most abundant Bacteria groups in most coleopterans and other insects are Proteobacteria and Firmicutes. This is not the case in Passalidae and Scarabaeidae, where Bacteroidetes and Firmicutes (both in larvae and adults) contain the most abundant OTUs, similar to mammal microbiomes[98]. Other wood feeding insects such as cockroaches and termites are colonized by different groups of bacteria. Bacteroidetes, Ellusimicrobia, Fibrobacter and Spirochaetas are among the main taxa dominating these insects.[34] These divergencies from the most common phylogenetic groups in insects suggests a certain degree of specialization by the beetle gut microbiota.

Larvae represent the most immature state of the insect's development, and their nutrient uptake will determine whether they will successfully reach adulthood. In addition to the main bacteria groups described above, larvae microbiomes from *Veturius sp.* have a high abundance of Archaea [51]. The presence of this group of microorganisms was confirmed in Passalid larvae and it was also found in both Scarabaeidae and Cerambycidae larvae samples (although less abundant). Previous work described the potential of archaeal MAGs in

Passalids to be methanogenic organisms[51]. On the other hand, Hackstein and Stumm analyzed methane production in arthropods and described how Scarabaeidae larvae have the capacity to produce this gas, which could be attributed to the archaeal OTUs present in the 16S rRNA analysis, as this could also be occurring in Passalid beetles [99].

Besides being enriched with archaeal genomes, larval gut microbiomes also had the most abundant core microbiome, with Proteobacteria as a main component. However, dominant groups vary among beetle families, for example Firmicutes is the dominant phylum of the Passalidae core microbiome; while Proteobacteria is dominant for Cerambycidae. This variation and the high number of OTUs in the Passalid larvae core microbiome could be explained by parental care and a consistent diet composed of pre-treated decomposing wood particles [36]. In the case of Cerambycidae, larvae feed directly on more unprocessed wood while Scarabaeidae has a more varied diet including soil hummus, decaying wood and decomposing organic matter[100,101]. Core microbiome from Passalid larvae includes Euryarchaeota OTUs, suggesting a probable role in the overall physiology of the beetle. The core microbiome of the larvae confirms the importance of several groups highlighted in this work, including the Ruminococcaceae family.

Larvae have more diverse core microbiomes than adult beetles, which are also dominated by Proteobacteria but in lower abundances This outcome is expected due to their more variable diet. Adults included in the survey have feeding behaviors that range from a variety of plants to wood in different stages of decomposition[37]. One of the main differences between adults from different samples is the presence of OTUs belonging to the Tenericutes phylum. Tenericutes is a group of intracellular bacteria found in many insects, highly abundant in the

Passalid microbiome, but not in the other two families. The possible role of this group is unknown as their main functions as gut residents are yet to be established [102]. Furthermore, the presence of OTUs from Tenericutes drives the differences in adults found in Passalidae core microbiomes compared to Sacarabaeidae, posing an interesting question for future research.

Finally, substrate is the most homogenous set of samples analyzed in this survey. The material is a combination of feces, chewed wood and leftovers from tunnel construction[36,100]. Thus, substrate is a mixture of environment microorganisms associated to wood decomposition and beetle gut microbiota [38]. Therefore, the higher richness observed in all substrate samples is probably a consequence of the constant disturbance upon the environment inside the log and the communities of bacteria and fungi that normally inhabit any decaying tree trunk in a forest. Differences could also be attributed to the oxidizing environmental conditions in the substrate when compared to an anaerobic digestive tract [50]. Under these conditions only aerobic and facultative anaerobic bacteria would be able to survive. On the other hand, the similarities amongst substrate samples are mostly due to high amounts of cellulose and plant material available for decomposition. Phyla present in every substrate sample have been reported to collaborate in decomposition of plant material in forest ecosystems. Members of the Proteobacteria, including *Phanerochaete velutina* and *Burkholderia* sp. seems to be involved in the decomposition of both sapwood and heartwood. Meanwhile, some actinobacteria like *Amycolatopsis sp*. are also capable of utilizing lignin as a sole carbon source [103]. All these microorganisms would be suppressed inside the beetle because of the environment limited by oxygen accessibility; favoring the abundance of anaerobic or facultative aerobic bacteria. On the other hand, the substrate microbiota, besides being responsible for a portion of cellulose degradation, is also involved in lignin degradation. Lignin degradation is an aerobic process, which would be impaired inside an intestinal tract with low concentrations of oxygen.

Following the evaluation of the microbial community structure of every sample type, I decided to analyze all three families of beetles separately. Differences in the ecology of each beetle family seem to influence the associated microbial community. All families chosen for the study commonly have a diet enriched with cellulose during their larvae stage. Meanwhile, adults have different feeding behaviors according to family [36,47,104]. Passalids, for example, have similar feeding behaviors between larvae and adults. Nonetheless, their microbiotas are different, which can be attributed to various factors. Nardi and colleagues described with detail the anatomy of the Passalidae gut in both developmental stages[105]. The larval digestive tract is less differentiated, even though it has the same sections as most insects (foregut, midgut, hindgut), only the midgut is well developed. Adults, on the other hand, have a more complex gut structure that can be divided into foregut (FG), midgut (MG), anterior hindgut (AHG) and posterior hindgut (PHG). The AHG structure occurs after a dramatic expansion during metamorphosis, and it is the area with less oxygen in the whole digestive tract [50]. Ruminococcaceae organisms are anaerobic, which could explain their ability to thrive in this environment. Thus, oxygen concentration across the gut should be kept at a minimum to sustain the relatively high abundances of anaerobic microorganisms, including Ruminococcaceae.

In this study, gut samples were not separated by sections, but rather examined as a whole. Hence, the true community structure of each individual compartment could be concealed by microbes from neighboring compartments. Adults have the most specialized and differentiated gut system. Therefore, there is a higher possibility of harboring specialized bacterial communities in each segment. In future research, it is recommended that different gut sections of both the adult and larvae are examined individually to better understand their microbiome composition and function.

Another factor that could explain differences in microbial populations amongst Passalid adult and larvae samples could be attributed to diet. Adults do not depend on substrate consumption as part of their nutrition and can survive out of raw wood [106]. On the contrary, during their larval stage, Passalids are not able to survive without the microbe-enriched substrate material [36]. The constant need for substrate material suggests that larvae require a vertical transfer of microbes from adult members of the same family group. One plausible explanation is that insects shed the exoskeletal lining of the foregut and hindgut each time they molt at their larval stage, disrupting most of the bacterial populations [33,107] and forcing continuous re-acquisition from the environment.

Because larvae are more sensitive to changes in their diet, it is possible that they will have a more stable and constant community structure. Alpha diversity analyses support this premise. A high evenness in larvae, as seen in Passalidae samples (Figure 10A), suggest a more stable community structure, being unlikely that any given individual in the population goes extinct[108]. Therefore, the established community structure needs to undergo continuous upkeeping. Resident microbes of the larvae must either survive all shed steps, or be restocked constantly by feeding, in order to achieve a steady microbiome. The results obtained in this work suggest that persistent ingestion of the substrate material (which is previously manipulated by the adult) inoculates microorganisms in the gut ecosystem of larvae allowing communities to be maintained.

Passalids and Scarabs (members of the family Scarabaeidae) are part of the same superfamily (Scarabaeidae), therefore similarities amongst them are expected[37]. However, both groups differ in terms of their behavior (Passalids are subsocial and Scarabs are not) and their feeding habits during the adult stage [48].Comparable to Passalids, Scarabaeidae larvae can thrive on soil hummus or decomposing wood, but in this case adults are mainly phytophagous[100]. Literature referring to Scarabaeidae beetles microbiome states that these organisms are dominated by phylogenetic groups with a fermentative metabolism such as Lactobacillales, Clostridiales, Bacillales, and Bacteroides, shown to be involved in the digestion of sugars [44]. During our survey, these groups were amongst the most abundant groups in larvae of both Passalidae and Sacarabaeidae. It is noteworthy that none of the previous studies have examined specifically the diversity of strict anaerobic bacteria, thus Ruminococcaceae is not mentioned in any of the previous reports describing the Scarabaeidae microbiome. This bacterial family, belonging to the Firmicutes, is abundant in both Scarabid and Passalid larvae; however, Passalid adults have a higher proportion of Ruminococcaceae than Scarabaeidae. Therefore, Ruminococcaceae organisms could be involved in cellulose degradation, particularly during their larval stage. However, evaluating this capability is out of the scope of this work.

Similar to the other two families, Cerambycidae larvae also inhabit large decomposing tree trunks in the forest where they feed for long periods of time. Once they undergo metamorphosis into adults, the adults can stay in a saproxylophagous or phytophagous diet

[104]. Bacterial groups associated with Cerambycidae have been evaluated in the past and these studies have shown similar results as the ones presented here, where Proteobacteria is the most abundant phylum in the gut [46]. Enterobacteriaceae, the most abundant family of OTUs found in larvae were also identified as the most abundant in other studies[109]. For example, Rizzi and colleagues performed DGGE to evaluate the uncultured microbiota and found four sequences from the phylum Firmicutes and one sequence from an uncultured Chitinophaga. I also found that Chitinophagaceae family was the most well represented family in the Cerambycidae substrate, but present also in adults and larvae.

Proteobacteria is also a dominant group in other xylophagous insects such as termites. Strategies for cellulose degradation in termites include the symbiosis with flagellates which code for cellulases or the production of their own endogenous enzymes [34]. Contrary to what occurs in Scarabaeidae and Passalidae, the Cerambycidae family, at least from the taxonomy profile standpoint, has more similarities with other herbivore insects than the other two beetle families. Lower termites and leaf cutting ants are associated with Proteobacteria (amongst other groups) that aid with carbohydrate metabolism. Further information is given by Johnson et al, as they describe how Cerambycidae beetles might have shifted from an acidic to an alkaline digestion[110]. This would explain the important abundances of Proteobacteria, as cellulolytic enzymes from this phylum have been described to work in alkaline conditions inside the termite gut. Finally, another possibility is that the synthesis of endogenous cellulose degrading enzymes from the host is helping the process, but to elucidate the potential cellulases of the three families of beetles, it would be necessary to sequence their whole genomes.

As previously described, the diversity of culturable microorganisms from these three families in Costa Rica has been already analyzed[47], obtaining hundreds of bacterial and fungal isolates from xylophagous beetles including Scarabaeidae, Cerambycidae and Passalidae, however, this study did not consider the unculturable microbial diversity. The results in this work employing culture-independent methods confirm that, despite their characteristic behavior, the gut community in Passalid larvae is more similar to other beetle larvae than to adults from its own species and even from the same decomposing log. Besides, a similar outcome was obtained for adults and substrate samples. Substrate samples, continuously exposed to the environment, should harbor a microbial community that resembles other environmental samples rather than the communities associated with insect hosts. However, among substrate samples, beetle family was the major trait predicting microbial populations. Such difference exist probably as a consequence of the constant inoculation with feces from each beetle family. The similarities in the microbial community composition among larvae and among adults regardless of beetle species could be explained by their comparable behavior and diet, specific to each developmental stage. The evidence above suggests that xylophagous beetles have similar strategies to metabolize their food and that their gut microbes are specific to each developmental stage. Further research comparing their gut microbiome to other non-xylophagous beetles is necessary to confirm these statements.

The survey with 16S rRNA amplicon libraries allowed a great number of samples to be analyzed at the same time. However, 16S rRNA sequencing has the limitation of being restricted to taxonomic data. Metagenomic data provides the opportunity to analyze the whole genomes of organisms in addition to elucidating their taxonomic assignment, including their enzymatic and metabolic capacities. One of the most abundant phylum found in the survey was Firmicutes, with Ruminococcaceae as its main representative. Organisms from the Ruminococcaceae family are known cellulose degraders in other systems, especially in cow rumen. Evidence of putative cellulases from Ruminococcaceae organisms in the metagenomic data indicate some of the insect hots cellulose degradation capacities can be attributed to this bacterial group. Previous work detected several types of cellulases in the metagenome of Passalid beetles from the genus *Veturius sp*, and generated several MAGs related to the Ruminococcaceae family [51]. This work expands on these previous results by creating a workflow that utilizes the available information to generate a comprehensive analysis of the mechanisms that these Ruminococcaceae organisms employ to degrade cellulose from a genomic perspective. While the available metagenomes were only from the Passalidae family, the same analysis developed in this work could be applied to the other beetle families as the data becomes available.

Ruminococcaceae GHs coding sequences are present in both larvae and adults but not in the substrate, therefore they seem to be favored by the anaerobic conditions in the gut. However, sequences from similar taxa such as *Clostridia* and order Clostridiales are present in all samples. Further, the proportion of genes in larvae and adults is similar, regardless of the small amount of sequences detected in the adult metagenome. A small number of sequences could have limited the resolution of the different types of GHs found in the adult as this particular metagenome has the lowest number of total genes. From these results, it seems that groups shown in Figure 15 are involved in carbohydrate metabolism in both larvae and adults. Previous studies regarding xylophagous beetles state that eukaryotic symbionts such as yeast and filamentous fungi are responsible for cellulose digestion[38,50,105], since these organisms exhibit enzymatic activity related to xylan, lignin and cellulose breakdown [38,47]. On the other hand, Ruminococcaceae sequences seem to make a small proportion of the carbohydrate metabolism genes detected within the system, which could be a result of the high amount of unclassified sequences but could also be due to the involvement of other bacteria in cellulose degradation. Ruminococcaceae sequences are a limited proportion of the carbohydrate metabolism within the system, but they include enzymes that target cellulose specifically. Previous research indicates that Ruminococcaceae organisms participate mostly in cellulose, cellobiose and glucose metabolism, while leaving the rest of carbohydrate substrates to other taxa for degradation [59,111].

Figure 16 shows the abundance of genes related exclusively to cellulose breakdown in each sample, including GHs and CBMs. Ruminococcaceae organisms seem to be involved in cellulose breakdown, but particularly in the larvae, coincident with the results obtained in the 16S rRNA gene analysis. The higher amount of GHs and CBMs sequences in substrate is probably a consequence of the combination of the normal microbiota driving wood decomposition and the vertical transmission of bacteria through their feces. The GHs found in high amounts in the substrate have been previously described in decaying wood, for example, Hori and colleagues report the presence of enzymes from the GH 6 and 9 families in fungi from the genus *Ganoderma, Flebia* and *Bjerkandera,* amongst others, in the white rot of trees. In this environment they contribute to cellulose, hemicellulose and pectin metabolism, but under aerobic conditions[112]. Meanwhile, the presence of only GH5 and GH9 in the microbiome of larvae indicates a different strategy for cellulose degradation in this developmental stage involving these two groups of enzymes.

A metagenomic search for GH related genes is a powerful tool, but it has the limitation that not every cellulase encountered within a Ruminococcaceae genome is going to be classified as such. If databases do not have enough information, sequences could be assigned to the most similar taxonomic group available instead of the actual taxon they belong to. On the other hand, this is not an issue when analyzing MAGs. The MAGs phylogeny construction relies on the alignment of certain ribosomal and housekeeping genes in the sccg collections and all the sequences will be associated to the closest taxonomy together, not each gene individually.

To further refine and assess the preliminary taxonomic identification assigned previously to the MAGs collection[49], a phylogenomic analysis was performed. Every MAG related to organisms in the Clostridiales Class was detected almost exclusively in larvae, however, this result could be influenced by the number of sequences in each metagenome. A recent study resolved a large quantity of rumen MAGs from almost 800gb of data, allowing the authors to identify only a small proportion of those MAGs to the species level[113]. Passalid metagenomes have much less data (approximately 1.5gb) and therefore it would be harder to have enough data to assemble most of the MAGs present in the samples. For example, only one MAG was identified in the phylogenomic tree as Lachnospiraceae. However, OTUs from this genus were more abundant in adults compared to larvae in the 16S rRNA analysis. Hence, the low yield of sequences in the adult metagenome, and also from the overall data probably influenced the diversity of MAGs retrieved in the analysis.

The Ruminococcaceae MAGs found in this study seem to be different from the *Ruminococcus* species described in rumen such as *R. albus* and *R. flavefaciens.* Meanwhile, some are related to other species such as *Subdoligranulum variable*, *Ruminococcus brommi* and *Clostridium sporosphaeroides sporosphaeroides* all of which have been described to be associated with the human gut microbiome [114–116]. Finally, there is a group of MAGs that did not cluster with any of the Ruminococcaceae reference genomes. These MAGs could represent new species, therefore it is important to attempt to culture them in the future to confirm this statement.

Following taxonomic assignment, a pangenomic analyses of the different Ruminococcaceae MAGs was performed. This analysis aimed to find putative function similarities in the different groups of Ruminococcaceae MAGs by defining protein clusters common to the entire clade. Few protein clusters related to carbon metabolism were shared, while none related to cellulose degradation was found in multiple MAGs. Even though none of the genes found in the pangenome were directly related to cellulose breakdown, oligosaccharide debranching enzymes such as alpha-1,6-glucosidase and glucan-1,4-alpha-glucosidase were found in both MAGs 519 and 174. These enzymes are probably related to the breakdown of glycogen for intracellular energy rather instead of the actual degradation of recalcitrant wood components.

Although the pangenomic analysis determined that protein clusters were not shared in each clade, several putative cellulase genes were found by analyzing MAGs individually. The presence of GH, CBMs, cohesin and dockering modules, all basic components of the cellulosomes, is strong evidence for the putative capacity of the Ruminococcaceae MAGs to metabolize cellulose crystals in the larval gut. Most members of the Ruminococcaceae family such as *R. flavefaciens, R. albus* and *Ruminococcus champanellensis* utilize both cellulosomes and free soluble cellulases[27,111]. The presence of GH5 and GH9 is noted in several *Ruminococcus* species within the literature[117]. GH5 corresponds to a family of beta-1,4 endoglucanases with a large variety of specificities[118] and it is the most abundant cellulase gene family detected in larvae according to the metagenomic analysis. Moreover, GH9 and GH48 are regarded as the major cellulose-degrading factors in bacteria, however, only GH9 was found in this study[119].

On the other hand, several MAGs also had CBMs domains. CBMs have no enzymatic activity, but are considered part of the cellulose degrading machinery, by promoting a prolonged interaction with the cellulosic substrate[23]. CBM 4 increases cellulose degrading activity in *Ruminococcus* sp. and can be found assembled to GH9 catalytic domains as part of the cellulosome [27]. CBM 6 has also been described to enhance the affinity for cellulose crystals, as part of a complex with GH5 [120].

MAGs ensembled from Bins 174, 503, 86 and 519 were chosen to be analyzed further because they encode for at least one known cellulase gene. In some cases, the cellulase enzymes are found together with scaffolding modules composed by several cohesin and dockerin domains[121]. GH9 in Bin 86 and Bin 503 were not found associated with cohesin and dockering modules. It is possible that these MAGs do not contain every sequence of the actual microorganism. On the other hand, both Bin 519 and Bin 174 showed gene clusters with enzyme and dockerin domains, even the presence of CBM in the case of Bin 174, supporting the presence of a cellulosome in these organisms. Furthermore, the location of these genes next to each other suggests co-expression or common regulation factors.

According to the RAST server, *Clostridium methylpentosum* (genome size: 3.5Mb) is the most closely related reference genome to both MAGs 174 and 519. *C. methylpentosum* is a strict anaerobe isolated from the human gut. It can catabolize pentoses as fermentable substrates, to produce compounds including acetate, propionate, n-propanol, CO_2, and H_2 [122].

On the other hand, MAG 86 and MAG 503 closest reference genomes were *Ethanoligenens harbinense* (genome size: 3.1Mb) and *Anaerotruncus colihominis* (genome size: 3.8Mb), respectively. Both bacteria are known H_2 producers and they are important carbohydrate fermenters in their respective environments [123,124].

Ruminococcaceae organisms are well known for their ability to metabolize several different carbohydrate molecules[121]. They are adapted to live both as free living organisms or as symbionts in the gut of different animals. Thus, they hold dense genomes, they can synthetize most of their own amino acids and ferment carbohydrates from different food sources. This is coherent with the amount of genes assigned to carbohydrate and amino acid metabolism in each MAG (Annex 6). These results suggest that Ruminococcaceae organisms probably metabolize cellobiose and glucose generated by decomposition of cellulose to provide fermentation substrates to synthesize pyruvate and then employ it for glycolysis and energy production. Byproducts from their metabolic activity could also be utilized by other microorganisms present in the gut. The higher abundance of Euryarchaeota in larvae from all three beetle families (although mainly in Passalidae), could be associated with methane production. Previous work demonstrated the presence of MAGs related to methane producing archaeas in the Passalid metagenomes[49]. Archaea such as *Methanomassiliicoccale*s and *Methanosarcina* are capable of producing methane by hydrogen dependent reduction of methanol and methylamines[34,125]. As mentioned above, all four reference genomes that were most closely related to Ruminococcaceae MAGs are H_2 producers[122–124]. Therefore, it is essential to prove experimentally that hydrogen is produced by Ruminococcaceae bacteria in the beetle gut to establish their role as intermediaries for methanogenesis.

Another interesting trait found through metabolic analysis of MAGs is the low abundance of genes related to the metabolism of aromatic compounds. Previous work concluded that Ruminococcaceae species are efficient degrading cellulose, but not the other recalcitrant compounds present in the plant cell wall [60,126]. Passalid beetles inhabit fallen logs with certain degree of decomposition, therefore the overall structure of the wood is already partially digested. Other microorganisms in the immediate surroundings such as the woody substrate had already started to decompose some of the more recalcitrant materials such as lignin and pectin. This could facilitate the acquisition of nutrients from wood, and increase accessibility to nutrients by Ruminococcaceae members, since recalcitrant substrates would be predigested already. Filamentous fungi and bacteria with laccase and lignin peroxidase activity have been described as residents of the Passalid beetle microbiota[47]. Hence, it should be possible for the beetle to breakdown aromatic compounds present in the wood, as long as aerobic conditions are available at some point along the digestive tract. Meanwhile, the strict anaerobes metabolize the leftover cellulose from the wood [50].

The results generated in this work suggest that Ruminococcaceae organisms could be providing nutrients to their host via production of acetate. The analysis summarized in Figure 22 shows that almost every enzyme needed to produce acetate was found in the selected MAGs. . Acetate is a SCFA that could be used as an energy source for the host, together with propionate and butyrate they are the main end products of carbohydrate breakdown [27,53,127]. In cows, volatile Fatty Acids are absorbed by the rumen epithelium and then transported via blood to the liver, where they are converted to other energy sources[128]. Thus, the information presented here suggests that Ruminococcaceae bacteria have the capacity to produce acetate inside Passalid larvae guts, which the insect could be using as an energy source.

Ruminococcaceae bacteria probably are performing their activity as components of a more complex consortium inside the gut, alongside other microorganisms, capable of complementing their enzymatic potential. Even though carbohydrate active enzymes in general and cellulases in particular were present in this bacterial family, a higher amount of related sequences were assigned to other bacterial groups. To achieve a full understanding of the fundamental interactions of the gut microbiome in Passalid beetles, other taxonomic groups should be further analyzed in depth, including Bacteroidetes or even other Firmicutes families.

As described in mammals, the gut microbiota of Passalids can be considered a bacterial organ [129]. Analysis of 16S amplicon within xylophagous beetle guts suggests that most of their metabolic capacity lies mostly on Firmicutes and Bacteroidetes phyla, at least for Scarabaeidae and Passalidae. Furthermore, the metagenomic data provided evidence on how most of the Passalid genes involved in processing recalcitrant carbohydrates are linked to anaerobic microorganisms.

Several authors conclude that microbes serving as symbionts in the intestinal tract of insects are generally involved in numerous functions, including nutrient uptake[7,97]. These is likely the case also for the microbiota inhabiting the gut of these beetles. For example, Passalidae larvae complete their development subsisting on a nutrient limited food source. One of the most important limitations of using wood as a food source is the lack of nitrogen in the diet. However, Passalid beetles have strategies to deal with this problem. Previous work found genetic evidence for nitrogen fixation in the Pasalid metagenomes[49]. Moreover, Ceja-

Navarro and colleagues demonstrated nitrogen fixation in this niche by isotopic analysis and detection of expression of the *nifH* gene[50]. Besides, we provide evidence of possible pathways through which microbes could enhance nutrient uptake by degrading cellulose and associated substrates. Many other insects such as termites and leaf-cutting ants rely on their microbiota to compensate for the lack of nutrients available in their main food source[4,34]. Passalid beetles are no exception, as they established symbiotic relations with commensal gut bacteria to address this issue.

To confirm most of the findings presented in this work, physiological and genetic characterizations of the Ruminococcaceae organisms are necessary. At the same time, the metabolic potential of Ruminococcaceae MAGs and the presence of putative cellulases make this bacterial group a target for future bioprospection. Therefore, these microorganisms should be isolated in pure culture. There are specific techniques for the isolation of rumen anaerobic microorganisms that employ cellulose enriched culture media. Hungate *et al* attempted to culture cellulolytic rumen bacteria using a broth that provided environmental and nutritional conditions simulating the rumen environment[130]. Subsequently, they developed a technique to isolate cellulolytic bacteria through a roll-tube method[131]. Nakamura and colleagues also described a method employing a six-well plate alongside the AnaeroPack System to isolate *Ruminococcus*[132].

A recent work by Ceja-Navarro and colleagues confirms some of the findings described in this thesis[49]. They employed multi-omics and integrated chemical analysis to describe how the anatomical structure and compartmentalization of the adult gut of Passalid beetles might select for different microbial communities along the digestive tract. This comprehensive work includes only adult samples and does not focus on any particular group of microorganisms. Fermentation, methanogenesis and acetogenesis seem to be enriched in the gut region were the gut wall is thickest (anterior hindgut). Also, in this region lower oxygen concentrations were found, providing the optimal conditions for Ruminococcaceae bacteria. They also found evidence of ethanol and lactate production, which are also present in the Ruminococcaceae MAGs described in this thesis.

Overall, our results confirm that xylophagous beetles have a complex and diverse microbiome colonizing their guts, as described in previous research. At the same time, their microbiome composition is driven mainly by sample type but is also strongly influenced by other factors. These analyses confirmed that bacteria in the Ruminococcaceae family are abundant members of the gut microbial communities in Passalid and Scarabaeidae beetles, thus deserving further research to elucidate their roles in the host ecophysiology.

8. Conclusions

Xylophagous beetle bacterial communities are mainly driven by the developmental stage, despite being from different geographic locations or a different family of beetles. However, both location and beetle family can also influence their gut microbiota.

Data suggest that larvae from Scarabaeidae and Passalidae families have a similar distribution of dominant groups of gut bacterial communities such as Firmicutes and Bacteroidetes. Such similarities suggest that both families have similar strategies to degrade cellulose.

The metagenomes of Passalidae insects encode for enzymes associated to cellulose degradation. A portion of these enzymes is related to the order Clostridiales and the Ruminococacaceae family. This information suggests an active role (but not exclusive) from this group in the overall cellulose and carbohydrate metabolism.

Larval metagenomes possess several MAGs affiliated to the Clostridiales class and the Ruminococcaceae family. These MAGs seem to correspond to both known and unknown species highlighting the Passalid gut environment as a source for possible enzymes.

Ruminococcaceae MAGs have the putative genes necessary to metabolize carbohydrates, amino acids and catabolize cellulose. Besides, some of these genes are clustered together, suggesting that they are being expressed simultaneously.

9. References

1. Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* **437**, 349–355 (2005).
2. Baumann, P. *et al. Genetics, Physiology, and Evolutionary Relationships of the Genus Buchnera: Intracellular Symbionts of Aphids*. *Annual Review of Microbiology* **49**,

(1995).

3. Hansen, A. K. & Moran, N. A. The impact of microbial symbionts on host plant utilization by herbivorous insects. *Mol. Ecol.* **23**, (2014).

4. Suen, G. *et al.* The Genome Sequence of the Leaf-Cutter Ant Atta cephalotes Reveals Insights into Its Obligate Symbiotic Lifestyle. *PLoS Genet.* **7**, e1002007 (2011).

5. Lodwig, E. M. *et al.* Amino-acid cycling drives nitrogen fixation in the legume–Rhizobium symbiosis. *Nature* **422**, 722–726 (2003).

6. Kaltenpoth, M., Göttler, W., Herzner, G. & Strohm, E. Symbiotic Bacteria Protect Wasp Larvae from Fungal Infestation. *Curr. Biol.* **15**, 475–479 (2005).

7. Dillon, R. J. & Dillon, V. M. The Gut Bacteria of Insects : Nonpathogenic Interactions. *Annu. Rev. Entomol.* **49**, 71–92 (2004).

8. Pinto-Tomás, A. A. *et al.* Symbiotic Nitrogen Fixation in the Fungus Gardens of Leaf-Cutter Ants. *Science (80-.).* **326**, (2009).

9. Aylward, F. O. *et al.* Metagenomic and metaproteomic insights into bacterial communities in leaf-cutter ant fungus gardens. *ISME J.* **6**, 1688–1701 (2012).

10. Auclair, J. L. Aphids: Their Biology, Natural Enemies and Control, World Crop Pests, Volume 2. *Int. J. Trop. Insect Sci.* **10**, 441 (1989).

11. Lamelas, A., Gosalbes, M. J., Moya, A. & Latorre, A. New Clues about the Evolutionary History of Metabolic Losses in Bacterial Endosymbionts, Provided by the Genome of Buchnera aphidicola from the Aphid Cinara tujafilina. *Appl. Environ. Microbiol.* **77**, 4446–4454 (2011).

12. Li, Z.-Q. *et al.* Character of cellulase activity in the guts of flagellate-free termites with different feeding habits. *J. Insect Sci.* **13**, 37 (2013).

13. York, W. S., Darvill, A. G., McNeil, M., Stevenson, T. T. & Albersheim, P. Isolation

and characterization of plant cell walls and cell wall components. *Methods Enzymol.* **118**, 3–40 (1986).

14. Gibson, L. J. The hierarchical structure and mechanics of plant materials. *J. R. Soc. Interface.* 2749–2766 (2012). doi:10.1098/rsif.2012.0341

15. Lebo, S. E. *et al.* Lignin. in *Kirk-Othmer Encyclopedia of Chemical Technology* (John Wiley & Sons, Inc., 2001). doi:10.1002/0471238961.12090714120914.a01.pub2

16. Klemm, D., Heublein, B., Fink, H.-P. & Bohn, A. Cellulose: fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Ed. Engl.* **44**, 3358–93 (2005).

17. Cousins, S. K. & Brown, R. M. Cellulose I microfibril assembly: computational molecular mechanics energy analysis favours bonding by van der Waals forces as the initial step in crystallization. *Polymer (Guildf).* **36**, 3885–3888 (1995).

18. Maruthamuthu, M., Jiménez, D. J., Stevens, P. & van Elsas, J. D. A multi-substrate approach for functional metagenomics-based screening for (hemi)cellulases in two wheat straw-degrading microbial consortia unveils novel thermoalkaliphilic enzymes. *BMC Genomics* **17**, 86 (2016).

19. Saini, J. K., Saini, R. & Tewari, L. Lignocellulosic agriculture wastes as biomass feedstocks for second-generation bioethanol production: concepts and recent developments. *3 Biotech* **5**, 337–353 (2015).

20. Bayer, E. A., Chanzy, H., Lamed, R. & Shoham, Y. Cellulose, cellulases and cellulosomes. *Curr. Opin. Struct. Biol.* **8**, 548–557 (1998).

21. Béguin, P. & Aubert, J.-P. The biological degradation of cellulose. *FEMS Microbiol Rev* **13**, 25–58 (1994).

22. Sharma, A., Tewari, R., Rana, S. S., Soni, R. & Soni, S. K. Cellulases: Classification, Methods of Determination and Industrial Applications. *Appl. Biochem. Biotechnol.*

179, 1346–1380 (2016).

23. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490-5 (2014).

24. Schwarz, W. H. The cellulosome and cellulose degradation by anaerobic bacteria. *Appl. Microbiol. Biotechnol.* **56**, 634–649 (2001).

25. Bayer, E. A., Lamed, R. & Himmel, M. E. The potential of cellulases and cellulosomes for cellulosic waste management. *Curr. Opin. Biotechnol.* **18**, 237–245 (2007).

26. Bayer, E. A., Shimon, L. J. W., Shoham, Y. & Lamed, R. Cellulosomes—Structure and Ultrastructure. *J. Struct. Biol.* **124**, 221–234 (1998).

27. Dassa, B. *et al.* Rumen Cellulosomics: Divergent Fiber-Degrading Strategies Revealed by Comparative Genome-Wide Analysis of Six Ruminococcal Strains. *PLoS One* **9**, e99221 (2014).

28. Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat. Rev. Microbiol.* **6**, 121–131 (2008).

29. Kobayashi, Y., Koike, S., Miyaji, M., Hata, H. & Tanaka, K. Hindgut microbes, fermentation and their seasonal variations in Hokkaido native horses compared to light horses. *Ecol. Res.* **21**, 285–291 (2006).

30. Hofmann, R. R. Evolutionary steps of ecophysiological adaptation and diversification of ruminants: a comparative view of their digestive system. *Oecologia* **78**, 443–457 (1989).

31. Pauli, J. N. *et al.* A syndrome of mutualism reinforces the lifestyle of a sloth. *Proc. R. Soc. B* **281**, 20133006 (2014).

32. Tokuda, G. *et al.* Cellulolytic environment in the midgut of the wood-feeding higher termite Nasutitermes takasagoensis. *J. Insect Physiol.* **58**, 147–54 (2012).

33. Engel, P. & Moran, N. A. The gut microbiota of insects - diversity in structure and function. *FEMS Microbiology Reviews* **37**, (2013).

34. Brune, A. & Dietrich, C. The Gut Microbiota of Termites: Digesting the Diversity in the Light of Ecology and Evolution. *Annu. Rev. Microbiol.* **69**, (2015).

35. Hanson, P. E. & Nishida, K. *Insects and other arthropods of tropical America*. (Cornell University Press, 2016).

36. Schuster, J. & Schuster, L. *Chapter 12: The evolution of social behavior in Passalidae (Coleoptera)*. *The Evolution of Social Behaviour in Insects and Arachnids* (Cambridge University Press, 1997).

37. Solís, A. *Costa Rica beetles : the most common families and subfamilies*. (Editorial INBio, 2002).

38. Urbina, H., Schuster, J. & Blackwell, M. The gut of Guatemalan passalid beetles: a habitat colonized by cellobiose- and xylose-fermenting yeasts. *Fungal Ecol.* **6**, 339–355 (2013).

39. Geib, S. M. *et al.* Microbial Community Profiling to Investigate Transmission of Bacteria Between Life Stages of the Wood-Boring Beetle, Anoplophora glabripennis. *Microb. Ecol.* **58**, 199–211 (2009).

40. Schuster, J. C. Odontotaenius floridanus New Species (Coleoptera: Passalidae): A Second U.S. Passalid Beetle. *Florida Entomol.* **77**, 474 (1994).

41. Reyes-Castillo, P. Passalidae: Morfologia y División en grandes Grupos; Géneros Americanos. *Folia Entomol. Mex.* (1970).

42. Schuster, J. C. & Schuster, L. B. Social Behavior in Passalid Beetles (Coleoptera:

Passalidae): Cooperative Brood Care. *Florida Entomol.* **68**, 266 (1985).

43. Warnecke, F. *et al.* Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560–565 (2007).

44. Egert, M., Wagner, B., Lemke, T., Brune, A. & Friedrich, M. W. Microbial Community Structure in Midgut and Hindgut of the Humus-Feeding Larva of Pachnoda ephippiata (Coleoptera: Scarabaeidae). *Appl. Environ. Microbiol.* **69**, 6659–6668 (2003).

45. Ayayee, P. *et al.* Gut Microbes Contribute to Nitrogen Provisioning in a Wood-Feeding Cerambycid. *Environ. Entomol.* **43**, (2014).

46. Schloss, P. D., Delalibera, I., Handelsman, J. & Raffa, K. F. Bacteria Associated with the Guts of Two Wood-Boring Beetles: Anoplophora glabripennis and Saperda vestita (Cerambycidae). **35**, 625–629 (2006).

47. Vargas-Asensio, G. *et al.* Uncovering the Cultivable Microbial Diversity of Costa Rican Beetles and Its Ability to Break Down Plant Cell Wall Components. *PLoS One* **9**, e113303 (2014).

48. Huang, S.-W., Zhang, H.-Y., Marshall, S. & Jackson, T. A. The scarab gut: A potential bioreactor for bio-fuel production. *Insect Sci.* **17**, 175–183 (2010).

49. Ceja-navarro, J. A. *et al.* Gut anatomical properties and microbial functional assembly promote lignocellulose deconstruction and colony subsistence of a wood-feeding beetle. *Nat. Microbiol.* (2019). doi:10.1038/s41564-019-0384-y

50. Ceja-Navarro, J. A. *et al.* Compartmentalized microbial composition, oxygen gradients and nitrogen fixation in the gut of Odontotaenius disjunctus. *ISME J.* **8**, 6–18 (2014).

51. Vargas-Asensio, G. V. Estructura, composición y potencial genético para degradar

celulosa de la microbiota intestinal del escarabajo. (Universidad de Costa Rica, 2019).

52. Shinkai, T., Ueki, T. & Kobayashi, Y. Detection and identification of rumen bacteria constituting a fibrolytic consortium dominated by Fibrobacter succinogenes. *Anim. Sci. J.* **81**, 72–79 (2010).

53. Piao, H. *et al.* Temporal dynamics of fibrolytic and methanogenic rumen microorganisms during in situ incubation of switchgrass determined by 16s rRNA gene profiling. *Front. Microbiol.* **5**, 1–11 (2014).

54. Mao, S., Zhang, M., Liu, J. & Zhu, W. Characterising the bacterial microbiota across the gastrointestinal tracts of dairy cattle: membership and potential function. *Sci. Rep.* **5**, 16116 (2015).

55. Hungate, R. E. *Rumen and Its Microbes.* (Elsevier Science, 1966).

56. Rainey, F. A. *Family VIII. Ruminococcaceae fam. nov. Bergey's Manual of Systematic Bacteriology* (2009).

57. Biddle, A., Stewart, L., Blanchard, J. & Leschine, S. Untangling the Genetic Basis of Fibrolytic Specialization by Lachnospiraceae and Ruminococcaceae in Diverse Gut Communities. *Diversity* **5**, 627–640 (2013).

58. Suen, G. *et al.* Complete genome of the cellulolytic ruminal bacterium Ruminococcus albus 7. *J. Bacteriol.* **193**, 5574–5 (2011).

59. Reau, A. J. La, Meier-Kolthoff, J. P. & Suen, G. Sequence-based analysis of the genus Ruminococcus resolves its phylogeny and reveals strong host association. *Microb. Genomics* **2**, (2016).

60. Christopherson, M. R. *et al.* Unique aspects of fiber degradation by the ruminal ethanologen Ruminococcus albus 7 revealed by physiological and transcriptomic analysis. *BMC Genomics* **15**, 1066 (2014).

61. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).

62. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

63. Edgar, R. C. & Bateman, A. Search and clustering orders of magnitude faster than BLAST. **26**, (2010).

64. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–72 (2006).

65. McMurdie, P. J., Holmes, S., Kindt, R., Legendre, P. & O'Hara, R. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).

66. Oksanen, J., Blanchet, F. & Kindt, R. Package 'vegan'. *Community Ecol.* (2013).

67. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

68. Markowitz, V. M. *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36**, D534-8 (2008).

69. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

70. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–86 (2007).

71. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

72. Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).

73. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics

data. *PeerJ* **3**, e1319 (2015).

74. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files Title. (2011).

75. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, (2015).

76. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).

77. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

78. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

79. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 1–8 (2018).

80. Wilkins, L. G. E., Ettinger, C. L., Jospin, G. & Eisen, J. A. Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia. *Sci. Rep.* **9**, 1–15 (2019).

81. Parks, D. H. *et al.* A proposal for a standardized bacterial taxonomy based on genome phylogeny. *bioRxiv* 256800 (2018). doi:10.1101/256800

82. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

83. Arkin, A. P. *et al.* The DOE Systems Biology Knowledgebase (KBase). *bioRxiv* 096354 (2016). doi:10.1101/096354

84. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* (2014). doi:10.1038/nmeth.3176

85. Van Dongen, S. & Abreu-Goodger, C. Using MCL to Extract Clusters from Networks. doi:10.1007/978-1-61779-361-5_15

86. Benedict, M. N., Henriksen, J. R., Metcalf, W. W., Whitaker, R. J. & Price, N. D. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* **15**, 8 (2014).

87. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

88. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation Yanbin. **40**, (2012).

89. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206-14 (2014).

90. Brettin, T. *et al.* RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5**, 8365 (2015).

91. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**, 75 (2008).

92. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

93. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2012).

94. Cazemier, A. E. *et al.* Fibre Digestion in Arthropods. 101–109 (1997).

doi:10.1016/S0300-9629(96)00443-4

95. Castillo, P. *et al.* Anomalously young volcanoes on old hot-spot traces: I. Geology and petrology of Cocos Island. *Geol. Soc. Am. Bull.* **100**, 1400–1414 (1988).

96. Yun, J.-H. *et al.* Insect gut bacterial diversity determined by environmental habitat, diet, developmental stage, and phylogeny of host. *Appl. Environ. Microbiol.* **80**, 5254–64 (2014).

97. Colman, D. R., Toolson, E. C. & Takacs-Vesbach, C. D. Do diet and taxonomy influence insect gut bacterial communities? *Mol. Ecol.* **21**, 5124–5137 (2012).

98. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science (80-.).* **320**, 1647–1651 (2008).

99. Hackstein, J. H. P. & Stumm, C. K. Methane production in terrestrial arthropods. *Microbiology* **91**, 5441–5445 (1994).

100. García-López, A., Micó, E., Numa, C. & Galante, E. Spatiotemporal Variation of Scarab Beetle Assemblages (Coleoptera: Scarabaeidae: Dynastinae, Melolonthinae, Rutelinae) in the Premontane Rain Forest in Costa Rica: A Question of Scale. *Ann. Entomol. Soc. Am.* **103**, 956–964 (2010).

101. Mason, C. J., Scully, E. D., Geib, S. M. & Hoover, K. Contrasting diets reveal metabolic plasticity in the tree-killing beetle, Anoplophora glabripennis (Cerambycidae: Lamiinae). *Sci. Rep.* **6**, (2016).

102. Brown, D. R. Phylum XVI. Tenericutes Murray 1984a, 356VP (Effective publication: Murray 1984b, 33.). in *Bergey's Manual® of Systematic Bacteriology* 567–723 (Springer New York, 2010). doi:10.1007/978-0-387-68572-4_5

103. Johnston, S. R., Boddy, L. & Weightman, A. J. Bacteria in decomposing wood and their interactions with wood-decay fungi. *FEMS Microbiol. Ecol.* **92**, fiw179 (2016).

104. Haack, R. A., Keena, M. A. & Eyre, D. *Life history and population dynamics of Cerambycidae. Chapter 2.* (1997).

105. Nardi, J. B. *et al.* Communities of microbes that inhabit the changing hindgut landscape of a subsocial beetle. *Arthropod Struct. Dev.* **35**, 57–68 (2006).

106. Pearse, A. S., Patterson, M. T., Rankin, J. S. & Wharton, G. W. The Ecology of Passalus Cornutus Fabricius, a beetle ehich lives in rotting logs. **6**, 455–490 (1936).

107. Hammer, T. J., Janzen, D. H., Hallwachs, W., Jaffe, S. P. & Fierer, N. Caterpillars lack a resident gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9641–9646 (2017).

108. Drobner, U., Bibby, J., Smith, B. & Wilson, J. B. The Relation between Community Biomass and Evenness: What Does Community Theory Predict, and Can These Predictions Be Tested? *Oikos* **82**, 295 (1998).

109. Rizzi, A. *et al.* Characterization of the bacterial community associated with larvae and adults of anoplophora chinensis collected in Italy by culture and culture-independent methods. *Biomed Res. Int.* **2013**, (2013).

110. Johnson, K. S. & Rabosky, D. Phylogenetic distribution of cysteine proteinases in beetles: evidence for an evolutionary shift to an alkaline digestive strategy in Cerambycidae. *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* **126**, 609–619 (2000).

111. Ben David, Y. *et al.* Ruminococcal cellulosome systems from rumen to human. *Environ. Microbiol.* **17**, 3407–3426 (2015).

112. Hori, C. *et al.* Genomewide analysis of polysaccharides degrading enzymes in 11 white- and brown-rot Polyporales provides insight into mechanisms of wood decay. *Mycologia* **105**, 1412–1427 (2013).

113. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic

sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).

114. Quentmeier, A. & Antranikian, G. Characterization of citrate lyase from Clostridium sporosphaeroides. *Arch. Microbiol.* **141**, 85–90 (1985).

115. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME J.* **6**, 1535–1543 (2012).

116. Holmstrøm, K., Collins, M. D., Møller, T., Falsen, E. & Lawson, P. A. Subdoligranulum variabile gen. nov., sp. nov. from human feces. *Anaerobe* **10**, 197–203 (2004).

117. Brumm, P. J. Bacterial genomes: what they teach us about cellulose degradation. *Biofuels* **4**, 669–681 (2013).

118. Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, H. & Henrissat, B. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* **12**, (2012).

119. Shi, W. *et al.* Comparative Genomic Analysis of the Endosymbionts of Herbivorous Insects Reveals Eco-Environmental Adaptations: Biotechnology Applications. *PLoS Genet.* **9**, (2013).

120. Bae, H.-J., Turcotte, G., Chamberland, H., Karita, S. & VÃ©zina, L.-P. A comparative study between an endoglucanase IV and its fused protein complex Cel5-CBM6. *FEMS Microbiol. Lett.* **227**, 175–181 (2003).

121. Ben David, Y. *et al.* Ruminococcal cellulosome systems from rumen to human. *Environ. Microbiol.* **17**, 3407–3426 (2015).

122. Himelbloom, B. H. & Canale-Parola, E. Clostridium methylpentosum sp. nov.: a ring-shaped intestinal bacterium that ferments only methylpentoses and pentoses. *Arch.*

Microbiol. **151**, 287–293 (1989).

123. Xing, D. *et al.* Ethanoligenens harbinense gen. nov., sp. nov., isolated from molasses wastewater. *Int. J. Syst. Evol. Microbiol.* **56**, 755–760 (2006).

124. Lawson, P. A. *et al.* Anaerotruncus colihominis gen. nov., sp. nov., from human faeces. *Int. J. Syst. Evol. Microbiol.* **54**, 413–417 (2004).

125. Brune, A. Methanogenesis in the Digestive Tracts of Insects. in *Handbook of Hydrocarbon and Lipid Microbiology* 707–728 (Springer Berlin Heidelberg, 2010). doi:10.1007/978-3-540-77587-4_56

126. La Reau, A. J. & Suen, G. The Ruminococci: key symbionts of the gut ecosystem. *J. Microbiol.* **56**, 199–208 (2018).

127. Henderson, G. *et al.* Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* **5**, (2015).

128. Moran, J. Chapter 5: How the rumen works. in *Tropical dairy farming : feeding management for small holder dairy farmers in the humid tropics* **73**, 71–87 (2005).

129. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science (80-.).* **307**, 1915–1920 (2005).

130. Hungate, R. E. Studies on cellulose fermentation. III. The culture and isolation of cellulose-decomposing bacteria from the rumen of cattle. *Growth (Lakeland)* 631–645 (1947).

131. Macy, J and Hungate, R. The Roll-Tube Method for Cultivation of Strict Anaerobes. *Bull. Ecol. Res. Comm.* **17**, 123–125 (1973).

132. Nakamura, K. *et al.* A Six-well Plate Method: Less Laborious and Effective Method for Cultivation of Obligate Anaerobic Microorganisms. *Microbes Environ.* **26**, 301–

306 (2011).